frontiers | Research Topics

Words in the world

Edited by
Gary Libben, Gonia Jarema, Juhani Järvikivi,
Eva Kehayia and Victor Kuperman

Published in
Frontiers in Communication
Frontiers in Psychology

ISSN 1664-8714
ISBN 978-2-8325-3582-0
DOI 10.3389/978-2-8325-3582-0

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Words in the world

Topic editors
Gary Libben — Brock University, Canada
Gonia Jarema — Montreal University, Canada
Juhani Järvikivi — University of Alberta, Canada
Eva Kehayia — McGill University, Canada
Victor Kuperman — McMaster University, Canada

Citation
Libben, G., Jarema, G., Järvikivi, J., Kehayia, E., Kuperman, V., eds. (2023). *Words in the world*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-3582-0

Table of contents

frontiers | Frontiers in Communication

TYPE Editorial
PUBLISHED 11 July 2023
DOI 10.3389/fcomm.2023.1234942

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Xiaolin Zhou,
Peking University, China

*CORRESPONDENCE
Gary Libben
gary.libben@gmail.com

RECEIVED 05 June 2023
ACCEPTED 12 June 2023
PUBLISHED 11 July 2023

CITATION
Libben G, Jarema G, Järvikivi J and Kehayia E (2023) Editorial: Words in the world. *Front. Commun.* 8:1234942.
doi: 10.3389/fcomm.2023.1234942

Editorial: Words in the world

Gary Libben[1]*, Gonia Jarema[2], Juhani Järvikivi[3] and Eva Kehayia[4]

[1]Department of Applied Linguistics, Brock University, St. Catharines, ON, Canada, [2]Department of Linguistics and Translation, Université de Montréal, Montreal, QC, Canada, [3]Department of Linguistics, University of Alberta, Edmonton, AB, Canada, [4]School of Physical and Occupational Therapy, McGill University, Montreal, QC, Canada

KEYWORDS

mental lexicon, lexical processing, dynamicity, individual differences, psycholinguistics

Editorial on the Research Topic
Words in the world

New ways to think about words

Words are at the core of our language and our lives. They constitute the means by which we share thoughts, understandings and feelings. They unite people and separate them. They are at once unique to each individual and the shared possession of a group. This Frontiers in Communication Research Topic addresses this complexity through research that is grounded in psycholinguistic experimentation and modeling. It presents evidence and insights concerning the nature of words across languages, their representations in the minds and brains of individuals, and their relations to people's environments and behaviors. This experimental psycholinguistic research, conducted across languages and across populations, suggests that the nature of words can only be understood as developing from language experience. Words are in the world—and so are our minds and brains.

This perspective is at the source of this Research Topic. It reflects the exciting developments that are taking place in the psycholinguistic study of language processing and in the cognitive sciences more generally. In many ways, these developments are entirely new. In other ways, they have their roots in foundational psycholinguistic research. It is noteworthy that it is now exactly 50 years since the publication of Derwing (1973) call for a behaviorally based approach to the study of language structure and language development. In that 1973 book, Derwing states that "a linguistic unit at any 'level' exists as a unit only because the language user treats it as a unit …as part of the language process of speech production and comprehension" (p. 305). This perspective highlights the fact that people's word knowledge is obligatorily linked to their language behavior in the world—their experience in understanding words, in producing them, and in relating them to each other and to ideas, entities, and emotions.

The inter-relationality of research on words in the world

Spanning languages

The comments above highlight the extent to which words are part of a highly integrated system that we typically call an individual's vocabulary. The characteristics of that system will differ considerably from language to language. It is not surprising, therefore, that in this Research Topic, a variety of languages are studied. These include Arabic, Chinese, Czech, Dutch, English, French, Greek, Italian, Norwegian, Spanish, and Swedish.

Words across the lifespan

Thinking about words as part of a highly integrated and dynamic vocabulary system brings us to the distinction between the vocabulary of a language and the vocabulary of an individual. The vocabulary of a language can change very dramatically over time. The vocabulary of an individual who speaks that language can also change very dramatically over their lifespan. The nature of age-related effects among people is investigated in four articles in this Research Topic. Garmann et al. examine the speech sounds produced by very young children and the extent to which those sound patterns are common across languages or conditioned by language-specific characteristics (in this case, Norwegian). The articles by Royle et al. and Montemurro et al. probe later stages in the lifespan. Royle et al. compared younger and older adults in terms of the extent to which they show links among French words that share features of spelling, semantics, and word structure. Montemurro et al. examined the naming ability of Italian-speaking older adults to uncover the role that cognitive reserve may play in the manifestation of age-related language processing phenomena. Finally, Sprenger et al. studied how aging may affect the ability of an individual to process idiomatic expressions.

Differences in the sematic and structural properties of words

The Sprenger et al. study offers an important link to another key issue in the Research Topic. This is the extent to which differences in the mental representation and processing of words are related to their semantic and structural properties. This matter is addressed in the article by Kulkova and Fischer who note that figurative expressions constitute a very significant part of language. The sematic differences among words are also addressed in the report of Wei and Gillon-Dowens, who investigated the extent to which differences between concrete and abstract words correspond to electrophysiological brain signals. They claim that different neural mechanisms may underlie non-attentive processing of abstract and concrete words.

As is evident in the articles in this Research Topic, the examination of semantic and structural differences among word types offers a way to advance research methodology and data analysis, enabling us to better understand how language is related to other aspects of cognitive function and moving us closer to a comprehensive psycholinguistics of word meaning and structure. The article by Baayen and Smolka presents a new means by which to model experimental findings that have distinguished the processing of German words from those of other languages. The study by de Almeida et al. uses an eye-tracking technique to understand how the processing of different verb types can help us better understand the relation between visual cognition and language processing. Marzi et al. offer a computational framework within which to understand how structural differences in words across languages can be modeled with a relatively small set of language independent principles.

The study by Spalding et al. brings our attention from semantic relations among words to the sematic relations within words. They concentrate on compound words such as *blackbird*, which can be seen as words that themselves contain words. These compound structures are also central to the study reported by Libben et al., which extends traditional single word experiments to the domain of multi-sentence texts by investigating the links among elements within compound words to other words in a text. The Libben et al. experiments involved the use of data from word and sentence typing. The typing technique is also central to the report of Feldman et al. who used it to uncover relations among word structure and variables such as word frequency and then linked those relations to models of cognitive and motor processing during typed production.

Words, health, and personal experience

Our use of words may be linked to our health both directly and indirectly. There are medical events such as stroke that can have direct and immediate effects on language processing ability. Often, however, as is the case in Alzheimer's disease, dementia, and other aspects of mental health, changes in language ability accompany other changes. This creates an opportunity to use language changes as an assessment tool and for symptom monitoring.

In the report by Havigerová et al., the written Czech production of persons experiencing depression was investigated across types of writing and in terms of orthographic variables and the structural characteristics of words. The authors report that their findings offer a technique that could be suitable for screening individuals at risk of depression. Manouilidou et al. investigated the processing of words among speakers of Greek with Mild Cognitive Impairment and with Alzheimer's Disease. They linked participants' cognitive abilities to their ability to process grammatical aspects of Greek words. In their meta-analytic review article, Kuzmina et al. surveyed patterns of language performance among bilingual persons who had experienced language difficulties as a result of damage to the brain. This study of bilingual aphasia revealed the important role played by patterns of language use in childhood.

Individuals' vocabularies that cross individual languages

The study of bilingual aphasia reported by Kuzmina et al. is related to a final matter addressed in this Research Topic. This is the simple but often overlooked fact that most people in the world speak more than one language. Thus, their mental store for words spans language vocabularies. The fact that this is the norm rather than the exception has enormous psycholinguistic consequences. The article by Cerni et al. reports on orthographic and phonological effects among native speakers of Italian who are learning English as a second language. In the study by Stephanich et al., the authors probe the theoretical challenges and potential solutions in the characterization of how bilingual individuals may switch languages within a single word. The study demonstrates how linking linguistic theory and language processing research can lead to important new insights.

Word research and the Frontiers in Communication

Psycholinguistic research on the representation and processing of words has progressed a great deal in recent years. New methodologies and approaches to experimentation and modeling have made it possible to study lexical processing across many more languages and within many more populations and contexts. In this way, we are moving to a more comprehensive understanding of *Words in the world*.

This greater understanding gives precedence to the highly integrated and relational nature of vocabulary knowledge. Our vocabularies are not simply storehouses of lexical knowledge. They are the dynamic systems that constitute our ability to understand and shape the world in which we live. Our lexical knowledge system grows and changes with us, so that it reflects many lifespan characteristics. Lexical knowledge systems may differ across groups and across social and cultural conditions. They are the leading edge of our language and, as such, define our cognitive reach and our ability to share our understandings with others. In this way, it may be that, as individuals, as groups, and as generations, word knowledge actually constitutes our *Frontiers in Communication*.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

This research was supported by the Social Sciences and Humanities Research Council of Canada Partnership Grant 895-2016-1008 ("*Words in the world*").

Acknowledgments

We wish to thank Laura Teddiman for her important contributions to this project and this Editorial Article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Derwing, B. L. (1973). *Transformational Grammar as a Theory of Language Acquisition: A Study in the Empirical Conceptual and Methodological Foundations of Contemporary Linguistics* (Vol. 10). Cambridge: Cambridge University Press.

ORIGINAL RESEARCH
published: 13 December 2018
doi: 10.3389/fpsyg.2018.02455

Written-Word Concreteness Effects in Non-attend Conditions: Evidence From Mismatch Responses and Cortical Oscillations

Dawei Wei[1,2] and Margaret Gillon-Dowens[2]*

[1] School of Foreign Languages, Lanzhou Jiaotong University, Lanzhou, China, [2] Cognitive Neuroscience of Language Laboratory, University of Nottingham Ningbo China, Ningbo, China

It has been widely reported that concrete words have processing advantages over abstract words in terms of speed and efficiency of processing, a phenomenon known as the concreteness effect. However, little is still known about the early time-course of processing concrete and abstract words and whether this concreteness effect can still persist in conditions where attention is not focused on the words presented (automatic processing). This study aimed to shed light on these issues by examining the electrophysiological brain responses to concrete and abstract words. While participants were engaged in a non-linguistic color tracking task presented in the center of the monitor screen, matched Chinese concrete and abstract single-character words appeared within a passive oddball paradigm, out of the focus of attention. In calculating visual Mismatch Negativity (vMMN), Event-related potentials (ERPs) to words of the same semantic category were compared when these words were presented as deviants and standards. Before 320 ms, both abstract and concrete words yielded vMMN with left-lateralized distribution, suggesting similar verbal processing at an initial processing stage. After 320 ms, only concrete words additionally elicited vMMN with a central distribution. Time frequency (TF) analysis of the results also revealed larger theta power increase (200–300 ms) and theta power phase locking (200–450 ms) for concrete than for abstract words. Interestingly, there was more alpha power decrease for abstract than for concrete words from 300 to 450 ms. This may reflect the greater difficulty in processing abstract meaning. Taken together, our ERP and TF results point to the existence of different neural mechanisms underlying non-attentive processing of abstract and concrete words.

Keywords: concreteness effect, visual mismatch negativity (vMMN), Chinese single-character words, time-frequency (TF) analysis, theta power increase, alpha power decrease, phase locking

OPEN ACCESS

Edited by:
Victor Kuperman, McMaster University, Canada

Reviewed by:
Lorenzo Magrassi, University of Pavia, Italy
Laura V. Sánchez-Vincitore, Universidad Iberoamericana, Dominican Republic

***Correspondence:**
Dawei Wei
da-wei.wei@outlook.com

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *15 November 2017*
Accepted: *20 November 2018*
Published: *13 December 2018*

Citation:
Wei D and Gillon-Dowens M (2018) Written-Word Concreteness Effects in Non-attend Conditions: Evidence From Mismatch Responses and Cortical Oscillations. Front. Psychol. 9:2455. doi: 10.3389/fpsyg.2018.02455

INTRODUCTION

The linguistic concreteness effect refers to the processing advantage of words representing imaginable, concrete concepts (e.g., apple) over those representing abstract concepts without sensory referents (e.g., skill). Specifically, in various experimental tasks such as lexical decision tasks, concrete words are often found to be recognized more quickly and more completely and remembered more accurately than abstract words (e.g., Strain et al., 1995; Peters and Daum, 2008; Barber et al., 2013). The concreteness effect has been of central interest among linguists and psychologists, as its sharp contrast between two semantic categories has been explored as a window onto the nature and structure of semantic representation in the human mind (Dalla Volta et al., 2014). Among the many semantic theories put forward to explain this effect, there are two predominant and competing ones, namely, the dual-coding theory (Paivio and Begg, 1971; Paivio, 1991) and the context availability theory (Schwanenflugel and Shoben, 1983; Schwanenflugel and Stowe, 1989). The dual coding theory proposes two qualitatively distinct systems of semantic processing, one verbal and the other non-verbal and imagery-based. Although both concrete and abstract words share the verbal processing system, concrete words can additionally engage the perceptual, imagery-focused system, which results in processing advantages. In comparison, the context availability theory argues for a common verbal system for both types of words but that concrete words can activate more associative information, resulting in a processing advantage over abstract words.

Based on clinical reports (Coltheart et al., 1980; Coslett and Saffran, 1989) and behavioral visual hemifield stimulation (Day, 1979; Chiarello et al., 1987), it has been argued that the verbal system is left-lateralized, while the imagery-based system is represented in the right hemisphere. So from a neural perspective, the dual coding theory predicts more involvement of right hemisphere for concrete than for abstract words. The context availability theory, in contrast, predicts that concrete words will have stronger activation than abstract words only in the left hemisphere. In recent decades, an increasing number of neuroimaging studies using functional Magnetic Resonance Imaging (fMRI) and Positron Emission Topography (PET) have investigated the neural substrates of concrete and abstract words. The findings from these studies are quite diverse. Some investigations have found more involvement of left hemisphere for concrete words (Jessen et al., 2000; Fiebach and Friederici, 2004), other studies have reported increased activity in the left hemisphere for abstract words, and bilateral activity for concrete words (Binder et al., 2005; Sabsevitz et al., 2005). Still others have described more activity only for abstract words (Kiehl et al., 1999; Perani et al., 1999; Noppeney and Price, 2004). In light of the inconsistent findings, it seems there is no clear consensus at present about the neural mechanisms of the concreteness effect.

In addition to metabolic neuroimaging studies, the event-related potential (ERP) technique has been widely used to explore this issue. One advantage of this technique is its millisecond-level time resolution, which is critical in uncovering details of when the cognitive processing of different stimulus events diverges. Findings from ERP studies can also be used to distinguish between various cognitive processes and representations based on the "spatial distinctiveness principle" (Holcomb et al., 1999). This principle assumes that heterogeneous cognitive processes tend to be associated with separate spatial distributions to a greater extent than will a single homogeneous process. Therefore, different topographical ERP patterns should be expected if concrete and abstract words are processed differently. Across different languages and a variety of tasks, such as explicit abstractness/imageability judgments and lexical decision, concrete words elicit larger N400 than abstract words, with a topographical distribution in posterior and anterior areas (Holcomb and Anderson, 1993; Holcomb et al., 1999; West and Holcomb, 2000; Tolentino and Tokowicz, 2009; Tsai et al., 2009). For example, using a lexical decision task, Zhang et al. (2006) investigated the neural dynamics of concrete and abstract Chinese words. Their findings replicated the typical concreteness N400 effect reported in other ERP studies (e.g., West and Holcomb, 2000). Concrete and abstract word effects also presented different scalp distributions, which was taken as supporting evidence for the dual-coding theory.

While most previous ERP studies have focused on the time interval of the N400 component, that is, 350–500 ms after stimulus onset, the dynamics of semantic processing at early latencies before 300 ms remain largely unexplored. Some recent studies, however, have provided evidence that semantic processing can be detected within the first 250 ms post-stimulus (Moscoso Del Prado Martín et al., 2006; Glenberg et al., 2008; Pulvermüller et al., 2009). Among the few ERP studies exploring early differences between abstract and concrete words, Wirth et al. (2008) investigated how a single-word context influenced the processing of abstract and concrete words. They manipulated concreteness (abstract vs. concrete words) and semantic relatedness (related vs. unrelated). In an early P1/N1 latency range, related as compared to unrelated abstract words were activated more in the left prefrontal cortex. For concrete words, this context effect was diminished, and abstract and concrete words performed differently according to the semantic context at a very early stage. Some other studies compared abstract and concrete words in terms of their elicitation of the recognition potential (RP) (Rudell, 1992) in a lexical decision task. RP is an early (before 300 ms) negative deflection sensitive to lexical-semantic manipulation (Hinojosa et al., 2004). In their study (Martiìn-Loeches et al., 2001), concrete words elicited larger RP than abstract words, but there were no differences in terms of the topographical distribution of RP. The authors interpreted this result as supporting a single unitary semantic system for both word types. Given that this result does not agree with previous N400 studies showing differences between abstract and concrete words, the authors suggested that, in their study, the similar processing mechanism for both word types may be confined to the early latency. Clearly, in order to have a complete picture of the differences between abstract and concrete word processing, it is necessary to find an approach which can focus on semantic processing at both early and late time windows. Another noteworthy point is that previous studies typically have used language-related tasks such as lexical decision and semantic categorization. With these tasks, participants need to pay active, conscious attention to the lexico-semantic stimuli in order to engage semantic analysis as expected. This, however, raises the question as to whether the N400 concreteness effect will still persist in non-attend conditions. A positive answer to this question can thus provide additional evidence for the concreteness effect, since the advantage would be generated by concrete words *per se* and independently of top-down task effects.

In fact, a few studies have found that semantic processing cannot only start early but can occur automatically, that is, without focused attention (Pulvermüller et al., 2005; Relander et al., 2009).

To answer these questions and fill the present research gaps in the area, the current study will focus on (visual) Mismatch Negativity (v)MMN as a critical ERP component in investigating the concreteness effect. MMN has long been demonstrated to be a reliable index of early and automatic processing of stimulus change (Näätänen, 2008). Its applications in spoken language studies are well documented (Pulvermüller and Shtyrov, 2006). Its early latency and sensitivity to semantic change (Shtyrov and Pulvermüller, 2007) thus make it an attractive tool for the current investigation into processing differences between abstract and concrete words. Numerous studies have found that the MMN component is sensitive to changes in speech features such as lexicality and semantics, as well as basic auditory attributes (e.g., acoustic duration, for a review see Pulvermüller and Shtyrov, 2006). For example, real words have been shown to elicit enhanced MMN compared to psycholinguistically matched pseudowords (e.g., Korpilahti et al., 2001; Kujala et al., 2007; Gu et al., 2012). Early semantic access has also been documented using an MMN approach. Shtyrov et al. (2004) compared MMN responses to two English verbs involving different body parts, the hand-related verb *pick* and leg-related verb *kick* in an oddball paradigm. Neuronal distribution of "pick" was found to be more lateral while that of "kick" more focal. This topographical difference seems to reflect different sensorimotor somatotopy rooted in the word semantics. In addition, recognition of the words started as early as 140–180 ms. The different activation patterns clearly demonstrate early, automatic semantic access to spoken words. These lexical and semantic MMN effects can be best explained by the long-term memory trace theory proposed by Pulvermüller and Shtyrov (2006). According to this theory, through associative language learning experience over time, long-term memory representations of existing words and their semantic features can be developed and shaped at the level of the neuronal network. These representations or traces over time become sufficiently robust to bolster rapid and automatic response to the word features, even in an attention-deprived condition (Shtyrov and Pulvermüller, 2007). Such responses can be neurophysiologically reflected as MMN effects (Pulvermüller and Shtyrov, 2006).

In recent years, MMN in the visual domain has also been revealed as an indicator of early visual change detection of elementary visual features such as color (Czigler et al., 2002), abstract sequential regularities (Stefanics et al., 2011) and even higher-level linguistic changes (Files et al., 2013; Shtyrov et al., 2013; Wang et al., 2013). In a passive oddball paradigm where matched Chinese single-character real words and pseudowords were compared, Wei et al. (2018) found that in native readers, only real words elicited vMMN as early as around 200 ms. In contrast, there was no sign of vMMN to either real words or pseudowords in a non-native control group with little learning experience in the Chinese language. This study largely replicated the lexical MMN findings in the auditory modality, thus supporting MMN as a modality-independent early index at least for lexical processing. Therefore, these studies suggest the potential role of vMMN in studying early linguistic effects. In addition, oddball paradigms, especially the identity oddball paradigm, where MMN is typically elicited, offer another advantage in controlling the physical variance of stimuli. In an identity oddball scheme, the role of deviant and standard stimuli in one sequence is swapped in the other one. The potential MMN is measured by comparing the same stimuli acting as deviants and standards across the two sequences. Therefore, it is possible to eliminate potential physical confounding and isolate a relatively pure cognitive process of interest (Pulvermüller and Shtyrov, 2006). In fact, physical variance between concrete and abstract words may have partially led to the ambiguous findings so far reported, so the identity paradigm could have potential for overcoming this confounding factor.

Additionally, to further explore the neural mechanisms of the concreteness effect, a time-frequency (TF) analysis of the acquired electroencephalography (EEG) data will also be carried out. While the "standard" ERP data analysis is to compute averaged evoked potentials time-locked to an event of interest and compare the amplitudes and latencies as a function of time across different experimental conditions, an increasingly popular development of EEG data analysis is to uncover their oscillatory neuronal dynamics in the frequency domain. Evoked activity features an identical phase ("phase-locked") and can be visible in averaged ERPs, whereas induced oscillations, though correlated with experimental conditions, have different onset times and/or phase jitter ("non-phase-locked"), and are therefore not visible in the time domain. However, many studies have demonstrated that both evoked and induced activities are important modes of brain functioning (Roach and Mathalon, 2008). The sum of both evoked and induced event-related oscillations is called overall spectral power or event-related spectral perturbation (ERSP). The phase-locking value/factor (PLV/PLF) or inter-trial (phase) coherence (ITC/ITPC) is another important complement to ERSP because it statistically assesses the event-related phase consistency of oscillations irrespective of their amplitude (Tallon-Baudry et al., 1997). PLV is between zero and one. If the phases during latency after the onset of a stimulus are a random distribution across trials, i.e., non-phase-locked activity, PLV would be zero. If, however, the phases are identical across all trials, i.e., strictly phase-locked activity, PLV would be one. In sum, time-frequency analyses of EEG signals provide additional information about which frequencies have the largest spectral power in a given latency and how their phases synchronize across time and space. These oscillatory dynamics have been found to be associated with various cognitive processes such as executive control, language and emotion, so time-frequency analysis can add new insights to the traditional ERP approach (Bastiaansen et al., 2012).

Neural oscillations are typically categorized into four frequency bands for analysis, delta (0–3 Hz), theta (4–7 Hz), alpha (8–12 Hz), beta (13–30 Hz, with the lower beta at 13–18 Hz), and gamma (above 30 Hz). A large corpus of studies over the last two decades has found a range of cognitive functions reflected in neuronal oscillatory dynamics in these four bands. For example, studies show that sustained gamma activity may play a critical role in holding working memory traces

(Kaiser et al., 2003; Jokisch and Jensen, 2007) and encoding of long-term memory representations (Osipova et al., 2006). In contrast, suppressing interference to prioritize task-relevant operations correlates with an increase in the oscillatory activity in the posterior alpha band (Klimesch et al., 1999; Jensen et al., 2002). Additionally, the theta frequency band has also been linked with memory retrieval and working memory (Bastiaansen et al., 2005, 2008). Delta band activity is involved in decision making and signal detection, to name but two functions (Başar et al., 2001; Yordanova et al., 2004). Of particular relevance to the current investigation are studies about the oscillatory correlates of lexical-semantic processing. Previous studies have found the importance of alpha band in semantic operations (Klimesch, 1999), and gamma band in semantic memory retrieval (Pulvermüller, 1999). Recent studies also seem to underscore the critical role of theta band oscillations in lexico-semantic processing. Bastiaansen et al. (2005) compared the oscillatory dynamics of meaning-bearing open-class words ("OC," e.g., nouns and verbs) and closed class words ("CC," e.g., determiners, articles), which carry more syntactic information, in a story-reading experiment. Both OC and CC words were found to have theta band power increase, as well as decrease in alpha and beta ranges. However, only OC words showed more theta synchronization over left temporal areas, which is argued to function in lexico-semantic retrieval (Indefrey and Cutler, 2004).

While quite a number of studies have investigated oscillatory patterns in language comprehension and general cognitive processing, only a handful of such studies have used MMN designs. Several TF studies of auditory MMN experiments have shown the important role of theta band in the generation of mismatch oscillatory responses (MOR) (Fuentemilla et al., 2008; Hsiao et al., 2009; Ko et al., 2012). In a visual MMN study on color change detection, it was found that theta band activity played a similar role in eliciting vMMN (Stothart and Kazanina, 2013). While no difference in alpha band power between deviant and standard stimuli has been reported in auditory MMN studies (e.g., Hsiao et al., 2009; Ko et al., 2012), two studies in the visual modality have indeed found such a difference. Stothart and Kazanina (2013) and Tugin et al. (2016) both described stronger alpha power decrease induced by deviants than standards. Tugin et al. (2016) also reported a larger increase of alpha band to deviant than standard items, despite an absence of vMMN in their study. Thus, given the confirmed roles of various frequency bands in linguistic processing as well as in domain-general cognition, a TF analysis of the ERP data of the current experiment will be carried out to paint a more complete picture of any concreteness-related effects.

In sum, the current study aims to further examine the effects of concreteness by exploring elicitation of the vMMN component, using a passive oddball paradigm. Based on previous findings, it is hypothesized that both concrete and abstract words may elicit vMMN, but concrete words may yield larger vMMN than abstract words. As previously stated, very few studies have attempted to explore the neuronal oscillatory dynamics of concrete and abstract words, so no specific hypothesis can be formulated about this issue. However, based on previous TF studies on language comprehension, especially the role of theta band spectral power and alpha power decrease in semantic processing (Klimesch et al., 1999; Bastiaansen et al., 2008, 2012), it is tentatively predicted that concrete words may elicit larger theta power increase than abstract words.

MATERIALS AND METHODS

Participants

Twenty-five neurologically and psychologically healthy college students (average age 21.5, *SD* = 1.2, male 8) participated in the experiment for course credit. They were all right-handed, native speakers of Chinese and had normal or corrected-to-normal vision and gave written informed consent in accordance with the Declaration of Helsinki before the experiment. The study was approved by the Research Ethics Committee of University of Nottingham Ningbo China and carried out in accordance with the approved regulations.

Stimuli

The stimuli included a set of five concrete and five abstract Chinese single-character words. These words were selected from a pool of 15 concrete and 15 abstract words with similar averaged ratings in number of strokes, frequency (counts per million), phonetic regularity and concreteness, which were chosen from the online Chinese Single-character Word Database (Liu et al., 2007). Twenty college students who did not take part in the ERP experiment were invited to carry out a rating task, judging the words in terms of emotional valence, arousal, frequency and concreteness on a seven-point Likert Scale (1 meaning the least positive, arousing, frequent or concrete). Based on the rating scores and data from the database, the final 10 words, five for each semantic category, were chosen. The final two sets are significantly different in concreteness ($p < 0.001$) but similar in all the other dimensions (all $ps > 0.1$) (see **Table 1** for a list of the stimuli. Numbers after the pinyin transcription represent the word tone).

Procedure

A complete trial is shown in **Figure 1**. During the experiment, participants were seated in a comfortable chair at a viewing distance of 50 cm from the monitor. Their task was to press keys when the black fixation cross in the centre of the screen changed to either red or green, that is, press "D" when the cross was green and "K" when the cross was red. They were asked to

TABLE 1 | Two different types of character stimuli.

Stimuli	1	2	3	4	5
Abstract	势	邦	宙	程	政
Pinyin	shi4	bang1	zhou4	cheng2	zheng4
Gloss	*trend*	*state*	*universe*	*procedure*	*politics*
Concrete	蚕	桃	岛	舌	绳
Pinyin	can2	tao2	dao3	she2	sheng2
Gloss	*silkworm*	*peach*	*island*	*tongue*	*rope*

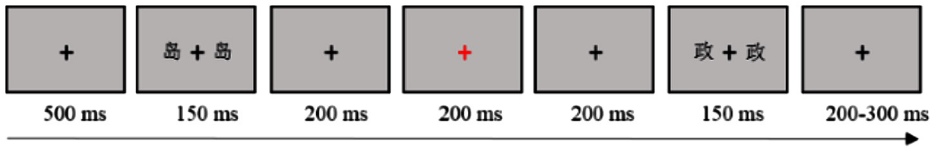


FIGURE 1 | Illustration of experimental procedure.

respond as quickly and accurately as possible. There were two blocks. The key-color correspondence was swapped in the second block. Each trial started with a black fixation cross for 500 ms on a gray background. Then two copies of a character (font: Sinsum; size: 75 pixels × 75 pixels; color: black; duration: 150 ms) were presented symmetrically to the right and to the left of the cross (i.e., perifoveally). Following this were three 200 ms-long blank screens, in the middle of which the fixation cross color was randomly changed to green or red. Participants had to respond to the color changes. This was followed by another two-character perifoveal presentation. Finally, there was a blank screen with a random duration of 200–300 ms. Throughout the trial, fixation crosses were always presented in the middle of the screen in order to maintain participants' attention to screen center and not to perifoveal stimuli.

EEG Data Recording

Brain Vision Recorder with a 32-channel EasyCap (Brain Products, Germany) (sampling rate 500 Hz, online band-pass filters at 0.01–100 Hz) was used to record electrophysiological (EEG) data. The reference electrode was attached to the tip of the nose. Horizontal and vertical electrooculography (EOG) were monitored using two electrodes placed on the outer canthi of the left eye and below the right eye, respectively. Impedances of all the electrodes (except EOGs which were below 10 kΩ), were maintained below 5 kΩ. In processing the data offline using Brain Vision Analyzer (Brain Products GmbH, Munich, Germany), a baseline of 100 ms prior to and 600 ms after the stimulus event was adopted. Butterworth Zero Phase digital shift (bandwidth 0.1–30 Hz, slope 24 dB/Octave), was used to filter data. Artifacts, including eye blinks, movement or muscle potentials, exceeding an absolute value of 100 µV at any electrode except EOGs were discarded. Four conditions of deviant concrete character, standard concrete character, deviant abstract character and standard abstract character were averaged, respectively, for further analysis. In the averaging procedure, the first three epochs in each block, and stimulus events preceded by a color fixation cross or by a button press were rejected. Data from four participants were discarded before proceeding to further statistical analysis, due to limited number of usable segments (<50% of all segments) in one of the four conditions, or low accuracy rate in the behavioral distraction task (<50%).

ERP Data Analysis

Choice of electrode sites for analysis was made with reference to previous literature on the semantic concreteness effect (West and Holcomb, 2000; Zhang et al., 2006) and on semantic MMN studies (Pulvermüller et al., 2001; Shtyrov et al., 2004). Midline and lateral site ERPs, quantified in respective areas by averaging the amplitudes of the electrodes included, were analyzed in ANOVAs of stimulus Type (deviant and standard), Condition (concrete and abstract) and Topographical factors. See **Figure 2** for the ERPs of deviants and standards of the two word classes. At the midline sites, the topographical factor was Region, including frontal (Fz, FC1, and FC2), central (Cz, CP1, and CP2) and parietal (Pz, P3, and P4) areas. At the lateral sites, the topographical factors included Hemisphere (left and right) and Region: frontal (F7/8), temporal (T7/8), parietal (P7/8), temporo-parietal (TP9/10) and occipital (O1/2) sites. Time windows spanning a 40 ms interval from 200 to 440 ms were selected for analysis, except the two windows of 320–360 ms and 360–400 ms which were merged. These windows were selected with references to previous studies (West and Holcomb, 2000; Zhang et al., 2006). The dependent variable is mean amplitude of voltages averaged across the electrodes in the selected regions and time windows. In all of the statistical analyses, Greenhouse-Geisser correction of the degrees of freedom was applied and the corrected *p*-values are reported, where appropriate.

Time-Frequency Analysis

To further uncover the oscillatory underpinnings of the mismatch responses, time-frequency analysis of the EEG data was conducted, also using Brain Vision Analyzer (Brain Products GmbH, Munich, Germany). Following the above-mentioned artifact rejection step in the EEG data processing, the four experimental conditions (Type: deviant and standard by Concreteness: concrete and abstract) were segmented with a baseline of 400 ms before the onset of stimuli. Here, the number of segments of standards was matched with that of deviants by selecting the standard stimuli preceding the deviant for analysis. This way, the potential influence of the number of the to-be-analyzed segments on TF representations, especially PLV, can be removed. Morlet wavelet transform was then applied to the segments with a frequency range from 1 to 40 Hz. In order to achieve a balance of time and frequency resolution (Tallon-Baudry, 2004), the Morlet parameter was set to 5. ERSPs that include both phase-locked and non-phase-locked oscillations were computed by averaging the absolute values of the wavelet transforms of segments, with a baseline correction based on the pre-stimulus interval from −400 ms to −200 ms (**Figure 3**). In calculating PLVs, the wavelet coefficients, instead of the spectral powers of the wavelet transform were measured first. PLVs were then computed and rectified before data extraction for statistical analysis (**Figure 4**).

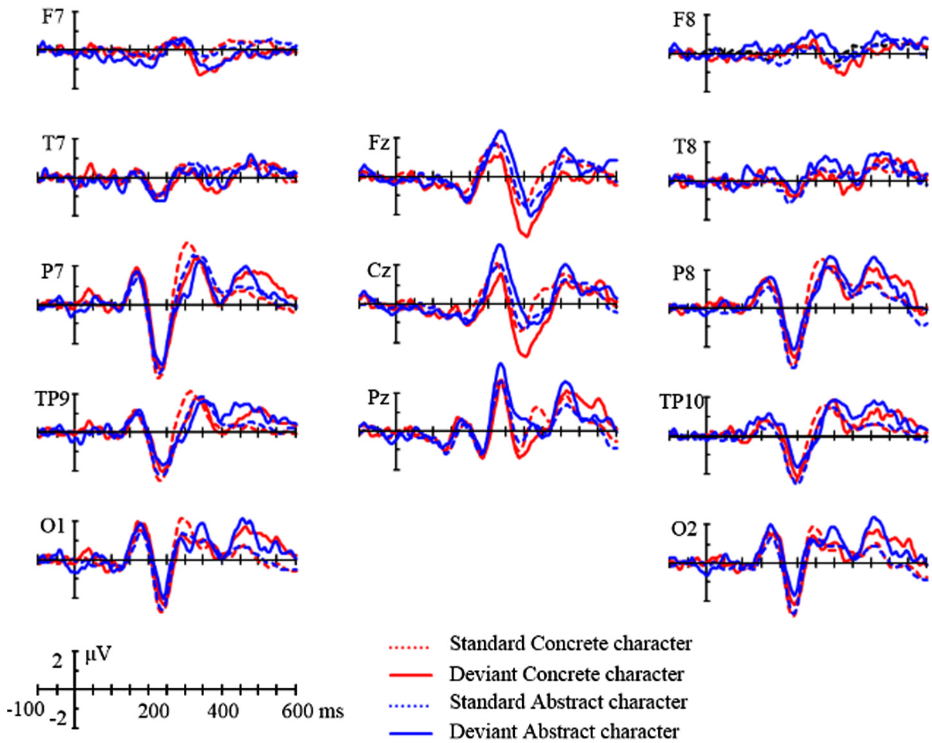


FIGURE 2 | Comparison of standard and deviant stimuli. ERP waveforms for standard (red dotted) and deviant (red solid) concrete character conditions and for standard (blue dotted) and deviant (blue solid) abstract character conditions at midline electrodes Fz, Cz, and Pz and lateral electrodes F7/8, T7/8, P7/8, TP9/10, and O1/2. Negativity is plotted downward.

RESULTS

Behavioral Performance

Paired samples *t*-test yielded no difference between the two blocks in terms of reaction times in response to the central fixation color changes, in the block with concrete deviants (572 ms) and in the other block with abstract deviants (581 ms), $t(20) = -1.32$, $p > 0.1$. In terms of accuracy, paired samples *t*-test suggested slightly higher accuracy rate in the block with concrete standards (93.2%) than in the other block with abstract standards (90.2%), $t(20) = -2.09$, $p < 0.05$.

ERP Results

As can be seen in **Figure 2**, both concrete and abstract words elicited canonical N/P1, N2 and P2 at most of the lateral sites. In addition, there is a noticeable divergence between concrete and abstract words when comparing their deviants and standards ERPs after 300 ms at midline sites. While for the concrete word deviants appeared to be more negative than standards over a time interval from 300 ms to around 450 ms, there is no such a sign for abstract words.

200–240 ms

At lateral sites, ANOVA yielded an interaction between Type, Hemisphere and Region, $(4,80) = 2.78$, $p = 0.05$, $\eta_p^2 = 0.12$. *Post hoc* analysis indicated that deviants were more negative than standards only at left frontal site, -0.38 vs. 0.05 μV, $p < 0.01$. At midline sites, ANOVA yielded an interaction between Type and Region, $F(2,40) = 6.46$, $p < 0.05$, $\eta_p^2 = 0.24$. Further analysis did not show difference between deviants and standards in any of the three regions, $ps > 0.1$.

240–280 ms

ANOVA at lateral sites indicated an interaction between Type and Hemisphere, $F(1,20) = 5.66$, $p < 0.05$, $\eta_p^2 = 0.22$. *Post hoc* analysis, however, did not indicate a difference between deviants and standards, $ps > 0.1$. At midline sites, there were no reliable effects.

280–320 ms

At lateral sites, there was an interaction of Type and Region, $F(4,80) = 10.36$, $p < 0.001$, $\eta_p^2 = 0.34$. *Post hoc* analysis indicated that deviants were more negative than standards at the temporo-parietal (0.27 vs. 1.04 μV, $p < 0.01$) and parietal (1.14 vs. 2.15 μV, $p < 0.01$) regions. There was an interaction of Type and Hemisphere, $F(20,1) = 5.55$, $p < 0.05$, $\eta_p^2 = 0.22$. *Post hoc* analysis indicated more negativity to deviants than to standards in the left hemisphere, 0.59 vs. 1.26, $p < 0.01$. At midline sites, there were no reliable effects.

320–400 ms

At lateral sites, ANOVA indicated an interaction of Type, Condition and Hemisphere, $F(1,20) = 9.09$, $p < 0.01$, $\eta_p^2 = 0.31$.

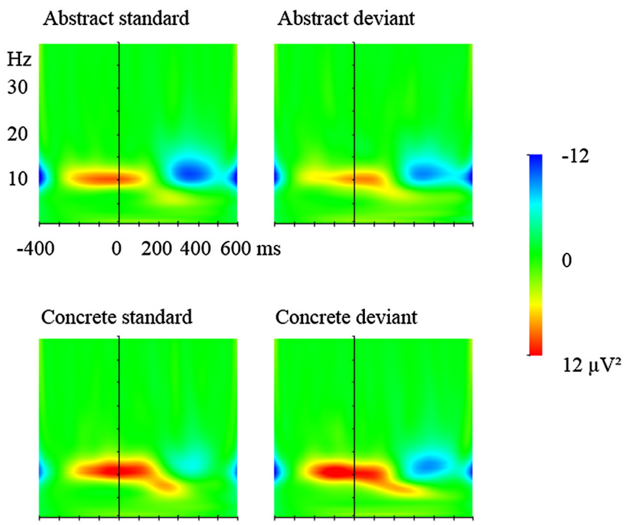


FIGURE 3 | Event-related spectral perturbation (ERSPs) of standard and deviant words. **(Upper)** Standard and deviant abstract words; **(lower)**: standard and deviant concrete words. The plots are based on electrode P7.

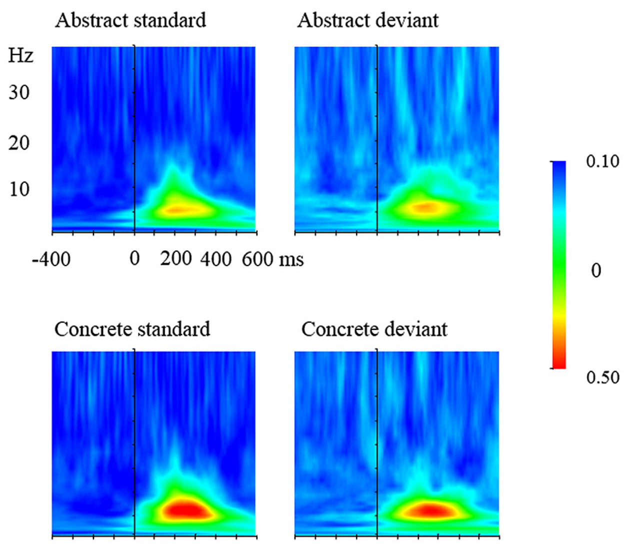


FIGURE 4 | Phase-locking value (PLVs) of standard and deviant words. **(Upper)** Standard and deviant abstract words; **(lower)** standard and deviant concrete words. The plots are based on electrode P7.

Post hoc analysis, however, did not show a difference between deviant and standard stimuli. At midline sites, there was an interaction of Type and Region, $F(2,40) = 5.12$, $p < 0.05$, $\eta_p^2 = 0.20$. Further analysis showed that deviants were more negative than standards in the frontal region, −1.60 vs. −0.82, $p < 0.05$. There was also an interaction of Type and Condition, $F(1,20) = 4.41$, $p < 0.05$, $\eta_p^2 = 0.18$. Further analysis indicated that deviants were more negative than standards only for concrete words, −1.33 vs. −0.01, $p < 0.05$.

400–440 ms

At lateral sites, ANOVA yielded an interaction of Type and Region, $F(4,80) = 10.35$, $p < 0.001$, $\eta_p^2 = 0.34$. *Post hoc* analysis indicated more negativity for deviants and standards in Frontal and Temporal regions, −0.22 vs. 0.26; 0.05 vs. 0.42, $ps < 0.05$. At midline sites, the interaction of Type, Condition and Region reached marginal significance, $F(2,40) = 3.77$, $p = 0.056$, $\eta_p^2 = 0.16$. Further analysis indicated that while deviants did not differ from standards for abstract words, $p > 0.1$, deviants were more negative than standards for concrete words in the frontal region, −0.51 vs. 0.84, $p < 0.05$.

Time-Frequency Results

Event-Related Spectral Perturbation (ERSP)

200–300 ms

Deviant and standard stimuli elicited an increase in overall spectral power, greatest at 6 Hz. In the lateral analysis, ANOVA yielded an interaction between Condition and Region, $F(4,80) = 10.41$, $p = 0.001$, $\eta_p^2 = 0.33$. *Post hoc* analysis revealed that concrete words produced more power increase than abstract words at the parietal and occipital sites, 5.16 vs. 3.83 μV^2, 3.03 vs. 1.40 μV^2, $ps < 0.05$. In the midline analysis, ANOVA yielded a main effect of Type, with deviant stimuli producing larger power increase than standard stimuli, $F(1,20) = 9.81$, $p < 0.01$, $\eta_p^2 = 0.33$, 3.64 vs. 1.90 μV^2. There was also an interaction of Condition and Region, $F(2,40) = 7.42$, $p < 0.01$, $\eta_p^2 = 0.27$. *Post hoc* analysis, however, did not show any differences between the two word classes.

300–450 ms

Both deviant and standard stimuli elicited a decrease in overall spectral power, greatest at 12 Hz (**Figure 3**). In the lateral analysis, ANOVA revealed an interaction between Type and Condition, $F(1,20) = 4.80$, $p < 0.05$, $\eta_p^2 = 0.19$. *Post hoc* analysis indicated that the difference between deviants and standards in concrete words reached marginal significance, $p = 0.062$, −2.58 vs. −0.95 μV^2, while there was no difference between deviant and standard abstract words, −2.37 vs. −3.28, $p > 0.1$. The interaction of Condition and Region reached significance, $F(4,80) = 4.16$, $p < 0.01$, $\eta_p^2 = 0.17$. *Post hoc* analysis showed that abstract words had stronger power decrease than concrete words in temporal-parietal (−2.01 vs. −0.97 μV^2, $p < 0.05$), parietal (−5.57 vs. −3.75 μV^2, $p < 0.05$), and occipital (−6.36 vs. −4.18 μV^2, $p = 0.058$) sites. In the midline analysis, no significant effects between standards and deviants were observed.

Phase-Locking Value (PLV)

Both deviant and standard stimuli elicited an increase in PLV, strongest at 4–8 Hz, between 150 and 450 ms (**Figure 4**). The increased PLV was focused in temporal and parietal areas. ANOVA was carried out separately for lateral and midline sites in the window of 200–450 ms. In the lateral analysis, ANOVA revealed an interaction between Type and Hemisphere, $F(1,20) = 5.92$, $p < 0.05$, $\eta_p^2 = 0.23$. *Post hoc* analysis found that deviants elicited higher PLV in the left hemisphere than in the right hemisphere, 0.29 vs. 0.26 μV. The interaction between Type and Region also reached significance, $F(4,80) = 5.87$,

p = 0.01, η_p^2 = 0.23. *Post hoc* analysis showed that deviants elicited higher PLV in frontal and temporal sites, 0.24 vs. 0.19 μV, 0.22 vs. 0.18 μV, $ps < 0.001$. ANOVA also revealed an interaction between Condition and Region, $F(4,80) = 5.22$, $p < 0.01$, η_p^2 = 0.21. *Post hoc* analysis showed that concrete words had higher PLV than abstract words at temporal, temporal-parietal, parietal and occipital sites, 0.21 vs. 0.19 μV, $p = 0.054$, 0.33 vs. 0.30 μV, $p < 0.01$, 0.36 vs. 0.32 μV, $p < 0.01$, 0.29 vs. 0.27 μV, $p < 0.05$. In the midline analysis, ANOVA revealed a main effect of Type, $F(1,20) = 12.27$, $p < 0.01$, η_p^2 = 0.38. *Post hoc* analysis showed that deviants elicited larger PLV than standards, 0.27 vs. 0.24 μV. No other effects reached significance.

DISCUSSION

The current experiment investigated early and automatic processing of concreteness in Chinese character reading. Concrete and abstract characters were presented perifoveally and participants were asked to carry out a non-linguistic distraction task presented in the middle of the screen. An oddball paradigm was used, where concrete and abstract characters acted as deviants and standards in one block and were swapped in the other block. Deviants elicited larger negative amplitudes than standards in various time windows across 200–440 ms. In early intervals before 320 ms, both concrete and abstract words elicited more negativity to deviants than to standards (i.e., vMMNs) only at lateral sites, specifically, at left frontal (200–240 ms), temporal-parietal and parietal sites (280–320 ms). At 320–400 ms, only concrete words yielded vMMN in the midline regions. At 400–440 ms, both categories of words elicited vMMNs at lateral frontal and temporal sites, but only concrete words tended to show vMMN at the central-frontal area.

As for the TF results, in the interval of 200–300 ms, deviants elicited larger theta power increase, peaking around 6 Hz in the central areas. Concrete characters elicited a larger theta power increase than abstract words at the lateral sites. Then from 300 to 450 ms, deviant stimuli elicited larger alpha power decrease than standard stimuli at the lateral sites. Abstract characters elicited larger alpha power decrease than concrete characters in lateral sites (specifically, parietal, temporo-parietal and occipital sites). In terms of PLV, in the interval of 200–450 ms, deviants elicited larger PLV than standards at the lateral (frontal and temporal sites) and central area; and concrete characters yielded larger PLV than concrete characters mainly at temporal, parietal and occipital sties. These results are discussed below.

Mismatch Responses of Concrete and Abstract Words

Before 320 ms, concrete and abstract word vMMNs had similar lateral distributions. According to the "spatial distinctiveness principle" (Holcomb et al., 1999), it can therefore be reasonably assumed that the two word categories have similar processing mechanisms. In these early windows, the negativities for both word classes are found to be left-lateralized, which may suggest that initially, verbal information is rapidly processed. Indeed, it has long been claimed that verbal systems are predominantly located in the left hemisphere (e.g., Coltheart et al., 1980; Chiarello et al., 1987; Kemmerer, 2014). Therefore, our data seem to be more in line with the dual-coding theory, which postulates a similar representational/verbal system for both concrete and abstract words.

After 320 ms, both concrete and abstract words yielded vMMNs at lateral frontal and temporal regions (400–440 ms). However, only concrete words additionally elicited vMMNs in the central area (320–400 ms) and middle frontal area (400–440 ms). This indicates that under non-attend conditions, concrete words still enjoy a semantic processing advantage, i.e., concreteness effect. This advantage in vMMN elicitation is arguably a result of the more perceptually salient referents of concrete words. These referents may contribute to a more robust and extensive neural network, and thus stronger memory representations than those of abstract words. Consequently, even in a context that discourages semantic processing, activation of concrete words can still be possible and is neurophysiologically reflected as the vMMN effect.

The concrete word vMMN with a frontal and central distribution is reminiscent of the well-documented concreteness N400 effects (Holcomb et al., 1999; West and Holcomb, 2000; Zhang et al., 2006; Kanske and Kotz, 2007; Tsai et al., 2009). The concreteness N400 effect is associated with more pronounced negativity over frontal sites, in contrast to the traditional semantic N400 effect characterized by centro-parietal morphology. Given the current non-attend oddball design, which does not encourage semantic processing, the emergent concreteness vMMNs suggest the resilient nature of the concreteness effect. In line with this, previous studies have also found concreteness N400 effects in tasks requiring various depths of semantic processing, such as image-generation (West and Holcomb, 2000), lexical decision (Tsai et al., 2009) or incidental memory retrieval (Nittono et al., 2002; Xiao et al., 2012). While the cortical generators of concrete words are strongly dependent on specific semantic categories (such as leg/arm/face-related words) (Hauk and Pulvermüller, 2004), some ERP studies on early semantic processing do find in source analyses that fronto-central areas are sensitive to form-related (Moscoso Del Prado Martín et al., 2006) and object-related (Moseley et al., 2013) words, which, to some extent, resemble the concrete words in the current study. Taken together, our data are consistent with previous studies and provide new evidence for distinct processing mechanisms responsible for abstract and concrete words. It should be noted, however, that possible parallels between the topographical distribution of ERP results and previous findings on semantic generators should be taken with caution due to the EEG inverse problem (Luck, 2005).

As discussed above, the similarities and differences from 200 to 440 ms between concrete and abstract words provide evidence for the dual-coding theory (Paivio, 1991). Our results therefore run against the context availability theory (Schwanenflugel and Shoben, 1983), according to which there should have been similar effects and scalp distributions for mismatch responses to both abstract and concrete words.

Early and Automatic Semantic Change Detection Revealed by vMMNs

In this experiment, words were contrasted in terms of their concreteness, but matched in visual complexity (stroke numbers), phonetic regularity, lexical frequency and emotionality (valence and arousal). Therefore, the ERP difference between deviant and standard stimuli is best attributed to the semantic contrast in terms of concreteness ratings. In other words, the vMMNs were elicited by rapid and automatic detection of semantic change. This finding extends previous vMMN investigations of non-linguistic object feature detection (for a recent review, see Kremláček et al., 2016) to the higher-order semantic level. In terms of the earliness of vMMN effects, both abstract and concrete words elicited vMMNs as early as 200–240 ms, which is also in line with previous studies on semantic processing in the auditory MMN field (Shtyrov et al., 2004; Pulvermüller et al., 2005; Shtyrov, 2010) as well as studies using other paradigms such as RP (Hinojosa et al., 2004) and lexical decision (Sereno and Rayner, 2003).

In addition, these early semantic vMMN effects occurred despite the non-attend design, and with a distracting non-linguistic color-tracking task, suggesting that the semantic processing was automatic in nature. Recently, Fujimura and Okanoya (2013), in one of the few studies on semantic processing using a vMMN design, explored automatic detection of the emotional connotations of Kanji words using an oddball paradigm similar to that of the current study. They also described early, enhanced vMMN to a strongly emotional deviant in comparison with neutral standards, at an early latency of 200–300 ms. In their study, however, the targets in the distraction task also served as members of the vMMN-eliciting emotion words. In addition, all the stimuli were centrally presented, despite the fact that emotional connotation is particularly attention-capturing, due to its relevance to survival (Lang et al., 1997). It could be argued, therefore, that their experimental setting may not have been stringent enough to avoid the attention-grabbing effects of the vMMN-eliciting critical stimuli. Thus, the automatic nature of those findings may be questionable. In the current study, with an improved design of perifoveal presentation of critical stimuli, vMMNs were, however, also observed, providing more compelling evidence for automatic semantic change detection.

The earliness and automaticity of semantic change detection in the current study may be explained by the long-term memory trace theory of linguistic MMN effects (Pulvermüller and Shtyrov, 2006; Shtyrov, 2010). As all the participants in the current experiment are native Chinese college students, they presumably have developed strong mental representations of the word stimuli tested here, or in other words, long-term memory representations of the linguistic items, as suggested and confirmed in previous MMN studies with spoken word stimuli (Pulvermüller et al., 2009). Additionally, the unique features of Chinese words may also play a role in the earliness of the concreteness effect. Chinese words are characterized by direct orthography-meaning correspondence with a relatively weak mediating role of phonology in accessing word meaning (Chen and Shu, 2001; Wang, 2011; Zhang et al., 2012, but see Perfetti and Tan, 1998). Indeed, there is no direct grapheme-phoneme correspondence in the writing system of Chinese, in contrast to alphabetic languages such as English where letters have more or less clear phonemic representations. Of particular relevance here, Zhang et al. (2006) investigated concreteness effects in Chinese words in a lexical decision task. Apart from the typical N400 concreteness effect at 300–500 ms, at an earlier interval of 200–300 ms, a concreteness effect was also observed. Therefore, the characteristics of Chinese writing in terms of semantic activation may play a part in these early semantic vMMN effects.

Oscillatory Characteristics

In the interval of 200–300 ms, both deviant and standard stimuli elicited overall spectral power increase in the theta range. This finding is in line with previous MMN studies in both auditory and visual modalities using elementary non-linguistic stimuli (Hsiao et al., 2009; Hsiao et al., 2010; Ko et al., 2012; Stothart and Kazanina, 2013), which have indicated the role of theta band oscillations in deviant detection, regardless of the mode of presentation. For example, in a vMMN study on frame color change detection, Stothart and Kazanina (2013) reported that similar theta power increase was yielded by both deviant and standard stimuli before 200 ms. In the current experiment, while at lateral sites the theta power between deviants and standards did not differ, deviants elicited larger theta power increase than standards in the central area. This difference between the two types of stimuli has seldom been reported in the above-mentioned MMN studies (Hsiao et al., 2009, 2010; Ko et al., 2012; Stothart and Kazanina, 2013). The reason for this may lie in the special stimulus category of language, compared to the basic-level auditory or visual features often targeted in those studies. Since deviant and standard words are different only in the dimension of semantic concreteness, the theta power difference may reflect such early semantic change detection. Consistent with this viewpoint, previous studies have indicated that theta band power increase can reflect lexical-semantic information retrieval (Bastiaansen et al., 2005, 2008). In fact, at parietal and occipital sites, the current experiment also found enhanced theta power in response to concrete words in comparison with abstract words, adding further evidence for the relevance of theta power to semantic processing.

In the following window of 300–450 ms, all the stimuli elicited alpha power decrease. Alpha power decrease has been reported in previous vMMN studies focusing on elementary visual feature detection (Stothart and Kazanina, 2013; Tugin et al., 2016). It has been suggested that while alpha power increase reflects task-related inhibition within a cortical area (Palva and Palva, 2007; Jensen et al., 2014), a decrease in the oscillatory amplitude suggests active neuronal processing (Pfurtscheller and Da Silva, 1999; Thut et al., 2012) and the enhanced attentional demands of processing (Klimesch, 1999; Bastiaansen et al., 2002;

Stothart and Kazanina, 2013; Wang and Bastiaansen, 2014). Therefore, the alpha power decrease here might indicate that both deviants and standards are undergoing further processing after the initial early-phase semantic processing. Stothart and Kazanina (2013) and Tugin et al. (2016) reported a larger alpha reduction for deviants than standards, whereas in the current study there is no such effect. One possible reason for this discrepancy might again lie in the types of stimuli used in the different studies. Though a non-attend paradigm is common to all three studies, linguistic items are relatively more difficult to process as deviants and standards, and thus demand more attentional resources than basic visual features such as color bars (Stothart and Kazanina, 2013) and moving dots (Tugin et al., 2016). Interestingly, in the current study, when deviant and standard abstract words are compared with deviant and standard concrete words, the former elicited a larger reduction in alpha band than the latter. This aligns well with previous studies showing that alpha power decrease, especially in its upper band (i.e., above 11 Hz) suggests semantic information retrieval, which demands enhanced attentional resources (Klimesch et al., 1997; Pérez et al., 2012). In line with this viewpoint, abstract words are considered to have fewer underlying semantic nodes in comparison with concrete words (Xiao et al., 2012), therefore making them less likely to support involuntary semantic retrieval in attention-deprived experimental conditions. An alternative explanation for the larger alpha power decrease for abstract words consists in the greater attentional demands for abstract word processing without involving semantic retrieval (Bastiaansen and Hagoort, 2006). However, due to the closely related association between attentional processing and semantic retrieval (Li and Ren, 2012), it is difficult to dissociate these two possibilities and the difference in alpha power decrease between abstract and concrete words may reflect a combination of both attentional demands and semantic processing. Further studies are needed to investigate the nature of alpha power decrease in response to factors of attention and semantics.

Relationship Between ERPs and TF Representations

While the TF data adds a new perspective to exploration of semantic change detection mechanisms, the temporal resolution of spectral power analysis is not as precise as that of ERP measurement, which means the onset and offset of the overall spectral power effects should be better taken as a rough estimation of relevant underlying neurophysiological activity (Clochon et al., 1996). Nevertheless, the spectral power latencies of theta and alpha bands largely overlap with the ERP effects in the current experiment. The initial theta power increase may correspond with the early automatic change detection as indexed by the early stage vMMN effects before 250 ms. The difference in the alpha power decrease effects between concrete and abstract words appears to overlap in timing with the difference in vMMN elicitation after 300 ms. However, it is the abstract words that yielded larger alpha power decrease compared to the concrete words. This points toward the notion that time-frequency analysis of the event-related EEG responses may characterize different neurophysiological mechanisms from traditional ERP analysis (Wang and Bastiaansen, 2014).

Phase Synchronization

Different from spectral power analysis, phase locking analysis characterizes the synchronization between spatially disparate areas into transitory neural networks, thus providing a unique tool to probe into neuronal dynamics. In the current study, the theta phase locking for deviants was larger than that for standards between 200 and 450 ms. This result coincides to some extent with the differences between the deviant and standard stimuli in the visual evoked potentials (i.e., vMMN effects). Therefore, theta phase locking is suggested to play a role in the vMMN effects. The result agrees well with a previous vMMN study on visual color bar detection (Stothart and Kazanina, 2013) as well as a series of auditory MMN studies (Fuentemilla et al., 2008; Hsiao et al., 2009; Bishop and Hardiman, 2010; Ko et al., 2012), indicating that the common role of theta phase locking in generating (v)MMN is independent of stimulus type and presentation modality. While these studies feature larger theta phase locking for deviant than standard stimuli in the right hemisphere, a clear left-lateralization was found in the current experiment. This topographic difference may be attributed to the processing of linguistic stimuli, in contrast to the basic feature perception explored in those studies. Similarly, larger phase locking for deviants than standards was found at frontal, temporal and midline sites, which appears to be consistent with the topographies of the vMMN effects, adding further support for the role of phase locking in language-related vMMNs in the time domain. In addition, in the window of 200–450 ms, concrete words elicited higher theta phase locking than abstract words in the midline analysis, which seems to be in line with the concrete word advantage in vMMN elicitation after 320 ms. This pattern may be attributable to the denser semantic links underlying concrete words, which neurophysiologically are represented as stronger functional connections. Therefore, the concreteness effect in the time domain is, to some extent, supported from the perspective of theta phase locking in the frequency domain.

CONCLUSION

The current study shows that both abstract and concrete words can be processed early and automatically as indexed by their elicited vMMNs. After 320 ms, a concreteness effect was observed, with only concrete words eliciting vMMNs in the middle areas, suggesting distinct processing routes for the two word types. Corroborating the concreteness effect in the time domain, concrete words elicited larger theta power increase and higher phase locking than did abstract words. Interestingly, abstract words yielded larger alpha power decrease in a later window after 300 ms, possibly resulting from the greater difficulty in processing abstract words in an attention-limited condition. Our study also shows the applicability of vMMN to semantic processing, especially at its early latency before 300 ms.

AUTHOR CONTRIBUTIONS

DW and MG-D conceived and designed the experiments. DW carried out experiments and analyzed data. MG-D contributed materials and recording tools. DW drafted the manuscript. MG-D gave critical revisions. DW and MG-D reviewed and approved the final manuscript.

ACKNOWLEDGMENTS

We would like to thank Alejandro Perez for his advice with data analysis and Joanna Yijin Lin for her help with data collection. This research work was supported by a Ph.D. scholarship granted to DW by the University of Nottingham Ningbo China (UNNC) and through a Research Support Scheme from Lanzhou Jiaotong University.

REFERENCES

Barber, H. A., Otten, L. J., Kousta, S.-T., and Vigliocco, G. (2013). Concreteness in word processing: ERP and behavioral effects in a lexical decision task. *Brain Lang.* 125, 47–53. doi: 10.1016/j.bandl.2013.01.005

Başar, E., Başar-Eroglu, C., Karakaş, S., and Schürmann, M. (2001). Gamma, alpha, delta, and theta oscillations govern cognitive processes. *Int. J. Psychophysiol.* 39, 241–248. doi: 10.1016/S0167-8760(00)00145-8

Bastiaansen, M., Böcker, K. B., and Brunia, C. H. (2002). ERD as an index of anticipatory attention? Effects of stimulus degradation. *Psychophysiology* 39, 16–28. doi: 10.1111/1469-8986.3910016

Bastiaansen, M., and Hagoort, P. (2006). Oscillatory neuronal dynamics during language comprehension. *Prog. Brain Res.* 159, 179. doi: 10.1016/S0079-6123(06)59012-0

Bastiaansen, M., Mazaheri, A., and Jensen, O. (2012). "Beyond ERPs: oscillatory neuronal dynamics," in *The Oxford Handbook of Event-Related Potential Components*, eds E. S. Kappenman and S. J. Luck (New York, NY: Oxford University Press), 31–50.

Bastiaansen, M. C., Oostenveld, R., Jensen, O., and Hagoort, P. (2008). I see what you mean: theta power increases are involved in the retrieval of lexical semantic information. *Brain Lang.* 106, 15–28. doi: 10.1016/j.bandl.2007.10.006

Bastiaansen, M. C., Van Der Linden, M., Ter Keurs, M., Dijkstra, T., and Hagoort, P. (2005). Theta responses are involved in lexical—Semantic retrieval during language processing. *J. Cogn. Neurosci.* 17, 530–541. doi: 10.1162/0898929053279469

Binder, J. R., Westbury, C. F., Mckiernan, K. A., Possing, E. T., and Medler, D. A. (2005). Distinct brain systems for processing concrete and abstract concepts. *J. Cogn. Neurosci.* 17, 905–917. doi: 10.1162/0898929054021102

Bishop, D. V. M., and Hardiman, M. J. (2010). Measurement of mismatch negativity in individuals: a study using single-trial analysis. *Psychophysiology* 47, 697–705. doi: 10.1111/j.1469-8986.2009.00970.x

Chen, H.-C., and Shu, H. (2001). Lexical activation during the recognition of Chinese characters: evidence against early phonological activation. *Psychon. Bull. Rev.* 8, 511–518. doi: 10.3758/BF03196186

Chiarello, C., Senehi, J., and Nuding, S. (1987). Semantic priming with abstract and concrete words: differential asymmetry may be postlexical. *Brain Lang.* 31, 43–60. doi: 10.1016/0093-934X(87)90060-5

Clochon, P., Fontbonne, J.-M., Lebrun, N., and Etévenon, P. (1996). A new method for quantifying EEG event-related desynchronization: amplitude envelope analysis. *Electroencephalogr. Clin. Neurophysiol.* 98, 126–129. doi: 10.1016/0013-4694(95)00192-1

Coltheart, M., Patterson, K., and Marshall, J. (1980). *Deep Dyslexia*. London: Routledge and Kegan Paul.

Coslett, H. B., and Saffran, E. M. (1989). Evidence for preserved reading in 'pure alexia'. *Brain* 112, 327–359. doi: 10.1093/brain/112.2.327

Czigler, I., Balazs, L., and Winkler, I. (2002). Memory-based detection of task-irrelevant visual changes. *Psychophysiology* 39, 869–873. doi: 10.1111/1469-8986.3960869

Dalla Volta, R., Fabbri-Destro, M., Gentilucci, M., and Avanzini, P. (2014). Spatiotemporal dynamics during processing of abstract and concrete verbs: an ERP study. *Neuropsychologia* 61, 163–174. doi: 10.1016/j.neuropsychologia.2014.06.019

Day, J. (1979). Visual half-field word recognition as a function of syntactic class and imageability. *Neuropsychologia* 17, 515–519. doi: 10.1016/0028-3932(79)90059-9

Fiebach, C. J., and Friederici, A. D. (2004). Processing concrete words: fMRI evidence against a specific right-hemisphere involvement. *Neuropsychologia* 42, 62–70. doi: 10.1016/S0028-3932(03)00145-3

Files, B. T., Auer, E. T., and Bernstein, L. E. (2013). The visual mismatch negativity elicited with visual speech stimuli. *Front. Hum. Neurosci.* 7:371. doi: 10.3389/fnhum.2013.00371

Fuentemilla, L., Marco-Pallarés, J., Münte, T., and Grau, C. (2008). Theta EEG oscillatory activity and auditory change detection. *Brain Res.* 1220, 93–101. doi: 10.1016/j.brainres.2007.07.079

Fujimura, T., and Okanoya, K. (2013). Event-related potentials elicited by pre-attentive emotional changes in temporal context. *PLoS One* 8:e63703. doi: 10.1371/journal.pone.0063703

Glenberg, A. M., Sato, M., and Cattaneo, L. (2008). Use-induced motor plasticity affects the processing of abstract and concrete language. *Curr. Biol.* 18, R290–R291. doi: 10.1016/j.cub.2008.02.036

Gu, F., Li, J., Wang, X., Hou, Q., Huang, Y., and Chen, L. (2012). Memory traces for tonal language words revealed by auditory event-related potentials. *Psychophysiology* 49, 1353–1360. doi: 10.1111/j.1469-8986.2012.01447.x

Hauk, O., and Pulvermüller, F. (2004). Neurophysiological distinction of action words in the fronto-central cortex. *Hum. Brain Mapp.* 21, 191–201. doi: 10.1002/hbm.10157

Hinojosa, J. A., Martin-Loeches, M., Muñoz, F., Casado, P., and Pozo, M. A. (2004). Electrophysiological evidence of automatic early semantic processing. *Brain Lang.* 88, 39–46. doi: 10.1016/S0093-934X(03)00158-5

Holcomb, P. J., and Anderson, J. E. (1993). Cross-modal semantic priming: a time-course analysis using event-related brain potentials. *Lang. Cogn. Process.* 8, 379–411. doi: 10.1080/01690969308407583

Holcomb, P. J., Kounios, J., Anderson, J. E., and West, W. C. (1999). Dual-coding, context-availability, and concreteness effects in sentence comprehension: an electrophysiological investigation. *J. Exp. Psychol. Learn. Mem. Cogn.* 25, 721–742. doi: 10.1037/0278-7393.25.3.721

Hsiao, F.-J., Cheng, C.-H., Liao, K.-K., and Lin, Y.-Y. (2010). Cortico-cortical phase synchrony in auditory mismatch processing. *Biol. Psychol.* 84, 336–345. doi: 10.1016/j.biopsycho.2010.03.019

Hsiao, F.-J., Wu, Z.-A., Ho, L.-T., and Lin, Y.-Y. (2009). Theta oscillation during auditory change detection: an MEG study. *Biol. Psychol.* 81, 58–66. doi: 10.1016/j.biopsycho.2009.01.007

Indefrey, P., and Cutler, A. (2004). "Prelexical and lexical processing in listening," in *The Cognitive Neurosciences III*, ed. M. S. Gazzaniga (Cambridge, MA: MIT Press), 759–774.

Jensen, O., Gelfand, J., Kounios, J., and Lisman, J. E. (2002). Oscillations in the alpha band (9–12 Hz) increase with memory load during retention in a short-term memory task. *Cereb. cortex* 12, 877–882. doi: 10.1093/cercor/12.8.877

Jensen, O., Gips, B., Bergmann, T. O., and Bonnefond, M. (2014). Temporal coding organized by coupled alpha and gamma oscillations prioritize visual processing. *Trends Neurosci.* 37, 357–369. doi: 10.1016/j.tins.2014.04.001

Jessen, F., Heun, R., Erb, M., Granath, D.-O., Klose, U., Papassotiropoulos, A., et al. (2000). The concreteness effect: evidence for dual coding and context availability. *Brain Lang.* 74, 103–112. doi: 10.1006/brln.2000.2340

Jokisch, D., and Jensen, O. (2007). Modulation of gamma and alpha activity during a working memory task engaging the dorsal or ventral stream. *J. Neurosci.* 27, 3244–3251. doi: 10.1523/JNEUROSCI.5399-06.2007

Kaiser, J., Ripper, B., Birbaumer, N., and Lutzenberger, W. (2003). Dynamics of gamma-band activity in human magnetoencephalogram during auditory pattern working memory. *Neuroimage* 20, 816–827. doi: 10.1016/S1053-8119(03)00350-1

Kanske, P., and Kotz, S. A. (2007). Concreteness in emotional words: ERP evidence from a hemifield study. *Brain Res.* 1148, 138–148. doi: 10.1016/j.brainres.2007.02.044

Kemmerer, D. (2014). *Cognitive Neuroscience of Language*. Hove: Psychology Press.

Kiehl, K. A., Liddle, P. F., Smith, A. M., Mendrek, A., Forster, B. B., and Hare, R. D. (1999). Neural pathways involved in the processing of concrete and abstract words. *Hum. Brain Mapp.* 7, 225–233. doi: 10.1002/(SICI)1097-0193(1999)7:4<225::AID-HBM1>3.0.CO;2-P

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Res. Rev.* 29, 169–195. doi: 10.1016/S0165-0173(98)00056-3

Klimesch, W., Doppelmayr, M., Pachinger, T., and Russegger, H. (1997). Event-related desynchronization in the alpha band and the processing of semantic information. *Cogn. Brain Res.* 6, 83–94. doi: 10.1016/S0926-6410(97)00018-9

Klimesch, W., Doppelmayr, M., Schwaiger, J., Auinger, P., and Winkler, T. (1999). "Paradoxical" alpha synchronization in a memory task. *Cogn. Brain Res.* 7, 493–501. doi: 10.1016/S0926-6410(98)00056-1

Ko, D., Kwon, S., Lee, G.-T., Im, C. H., Kim, K. H., and Jung, K.-Y. (2012). Theta oscillation related to the auditory discrimination process in mismatch negativity: oddball versus control paradigm. *J. Clin. Neurol.* 8, 35–42. doi: 10.3988/jcn.2012.8.1.35

Korpilahti, P., Krause, C. M., Holopainen, I., and Lang, A. H. (2001). Early and late mismatch negativity elicited by words and speech-like stimuli in children. *Brain Lang.* 76, 332–339. doi: 10.1006/brln.2000.2426

Kremláček, J., Kreegipuu, K., Tales, A., Astikainen, P., Põldver, N., Näätänen, R., et al. (2016). Visual mismatch negativity (vMMN): a review and meta-analysis of studies in psychiatric and neurological disorders. *Cortex* 80, 76–112. doi: 10.1016/j.cortex.2016.03.017

Kujala, T., Tervaniemi, M., and Schröger, E. (2007). The mismatch negativity in cognitive and clinical neuroscience: theoretical and methodological considerations. *Biol. Psychol.* 74, 1–19. doi: 10.1016/j.biopsycho.2006.06.001

Lang, P. J., Bradley, M. M., and Cuthbert, B. N. (1997). "Motivated attention: affect, activation, and action," in *Attention and Orienting: Sensory and Motivational Processes*, eds P. J. Lang, R. F. Simons, and M. T. Balaban (Mahwah, NJ: Lawrence Erlbaum Associates Publishers), 97–135.

Li, X., and Ren, G. (2012). How and when accentuation influences temporally selective attention and subsequent semantic processing during on-line spoken language comprehension: an ERP study. *Neuropsychologia* 50, 1882–1894. doi: 10.1016/j.neuropsychologia.2012.04.013

Liu, Y., Shu, H., and Li, P. (2007). Word naming and psycholinguistic norms: Chinese. *Behav. Res. Methods* 39, 192–198. doi: 10.3758/bf03193147

Luck, S. J. (2005). *An Introduction to the Event-Related Potential Technique*. Cambridge, MA: MIT Press.

Martiìn-Loeches, M., Hinojosa, J. A., Fernández-FrìAs, C., and Rubia, F. J. (2001). Functional differences in the semantic processing of concrete and abstract words. *Neuropsychologia* 39, 1086–1096. doi: 10.1016/S0028-3932(01)00033-1

Moscoso Del Prado Martín, F., Hauk, O., and Pulvermüller, F. (2006). Category specificity in the processing of color-related and form-related words: an ERP study. *Neuroimage* 29, 29–37. doi: 10.1016/j.neuroimage.2005.07.055

Moseley, R. L., Pulvermüller, F., and Shtyrov, Y. (2013). Sensorimotor semantics on the spot: brain activity dissociates between conceptual categories within 150 ms. *Sci. Rep.* 3:1928. doi: 10.1038/srep01928

Näätänen, R. (2008). Mismatch negativity (MMN) as an index of central auditory system plasticity. *Int. J. Audiol.* 47, S16–S20. doi: 10.1080/14992020802340116

Nittono, H., Suehiro, M., and Hori, T. (2002). Word imageability and N400 in an incidental memory paradigm. *Int. J. Psychophysiol.* 44, 219–229. doi: 10.1016/S0167-8760(02)00002-8

Noppeney, U., and Price, C. J. (2004). Retrieval of abstract semantics. *Neuroimage* 22, 164–170. doi: 10.1016/j.neuroimage.2003.12.010

Osipova, D., Takashima, A., Oostenveld, R., Fernández, G., Maris, E., and Jensen, O. (2006). Theta and gamma oscillations predict encoding and retrieval of declarative memory. *J. Neurosci.* 26, 7523–7531. doi: 10.1523/JNEUROSCI.1948-06.2006

Paivio, A. (1991). Dual coding theory: retrospect and current status. *Can. J. Psychol.* 45, 255–287. doi: 10.1037/h0084295

Paivio, A., and Begg, I. (1971). Imagery and associative overlap in short-term memory. *J. Exp. Psychol.* 89, 40–45. doi: 10.1037/h0031193

Palva, S., and Palva, J. M. (2007). New vistas for α-frequency band oscillations. *Trends Neurosci.* 30, 150–158. doi: 10.1016/j.tins.2007.02.001

Perani, D., Cappa, S. F., Schnur, T., Tettamanti, M., Collina, S., Rosa, M. M., et al. (1999). The neural correlates of verb and noun processing: a PET study. *Brain* 122, 2337–2344. doi: 10.1093/brain/122.12.2337

Pérez, A., Molinaro, N., Mancini, S., Barraza, P., and Carreiras, M. (2012). Oscillatory dynamics related to the Unagreement pattern in Spanish. *Neuropsychologia* 50, 2584–2597. doi: 10.1016/j.neuropsychologia.2012.07.009

Perfetti, C. A., and Tan, L. H. (1998). The time course of graphic, phonological, and semantic activation in Chinese character identification. *J. Exp. Psychol. Learn. Mem. Cogn.* 24, 101–118. doi: 10.1037//0278-7393.24.1.101

Peters, J., and Daum, I. (2008). Differential effects of normal aging on recollection of concrete and abstract words. *Neuropsychology* 22:255. doi: 10.1037/0894-4105.22.2.255

Pfurtscheller, G., and Da Silva, F. L. (1999). Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clin. Neurophysiol.* 110, 1842–1857. doi: 10.1016/S1388-2457(99)00141-8

Pulvermüller, F. (1999). Words in the brain's language. *Behav. Brain Sci.* 22, 253–279. doi: 10.1017/S0140525X9900182X

Pulvermüller, F., Assadollahi, R., and Elbert, T. (2001). Neuromagnetic evidence for early semantic access in word recognition. *Eur. J. Neurosci.* 13, 201–205. doi: 10.1046/j.0953-816X.2000.01380.x

Pulvermüller, F., and Shtyrov, Y. (2006). Language outside the focus of attention: the mismatch negativity as a tool for studying higher cognitive processes. *Prog. Neurobiol.* 79, 49–71. doi: 10.1016/j.pneurobio.2006.04.004

Pulvermüller, F., Shtyrov, Y., and Hauk, O. (2009). Understanding in an instant: neurophysiological evidence for mechanistic language circuits in the brain. *Brain Lang.* 110, 81–94. doi: 10.1016/j.bandl.2008.12.001

Pulvermüller, F., Shtyrov, Y., and Ilmoniemi, R. (2005). Brain signatures of meaning access in action word recognition. *J. Cogn. Neurosci.* 17, 884–892. doi: 10.1162/0898929054021111

Relander, K., Rämä, P., and Kujala, T. (2009). Word semantics is processed even without attentional effort. *J. Cogn. Neurosci.* 21, 1511–1522. doi: 10.1162/jocn.2009.21127

Roach, B. J., and Mathalon, D. H. (2008). Event-related EEG time-frequency analysis: an overview of measures and an analysis of early gamma band phase locking in schizophrenia. *Schizophr. Bull.* 34, 907–926. doi: 10.1016/j.neuroimage.2005.04.012

Rudell, A. (1992). Rapid stream stimulation and the recognition potential. *Electroencephalogr. Clin. Neurophysiol.* 83, 77–82. doi: 10.1016/0013-4694(92)90135-5

Sabsevitz, D. S., Medler, D. A., Seidenberg, M., and Binder, J. R. (2005). Modulation of the semantic system by word imageability. *Neuroimage* 27, 188–200. doi: 10.1016/j.neuroimage.2005.04.012

Schwanenflugel, P. J., and Shoben, E. J. (1983). Differential context effects in the comprehension of abstract and concrete verbal materials. *J. Exp. Psychol. Learn. Mem. Cogn.* 9, 82–102. doi: 10.1037/0278-7393.9.1.82

Schwanenflugel, P. J., and Stowe, R. W. (1989). Context availability and the processing of abstract and concrete words in sentences. *Read. Res. Q.* 24, 114–126. doi: 10.2307/748013

Sereno, S. C., and Rayner, K. (2003). Measuring word recognition in reading: eye movements and event-related potentials. *Trends Cogn. Sci.* 7, 489–493. doi: 10.1016/j.tics.2003.09.010

Shtyrov, Y. (2010). Automaticity and attentional control in spoken language processing: neurophysiological evidence. *Ment. Lexicon* 5, 255–276. doi: 10.1075/ml.5.2.06sht

Shtyrov, Y., Goryainova, G., Tugin, S., Ossadtchi, A., and Shestakova, A. (2013). Automatic processing of unattended lexical information in visual oddball presentation: neurophysiological evidence. *Front. Hum. Neurosci.* 7:421. doi: 10.3389/fnhum.2013.00421

Shtyrov, Y., Hauk, O., and Pulvermüller, F. (2004). Distributed neuronal networks for encoding category-specific semantic information: the mismatch negativity to action words. *Eur. J. Neurosci.* 19, 1083–1092. doi: 10.1111/j.1460-9568.2004.03126.x

Shtyrov, Y., and Pulvermüller, F. (2007). Language in the mismatch negativity design: motivations, benefits, and prospects. *J. Psychophysiol.* 21, 176–187. doi: 10.1027/0269-8803.21.34.176

Stefanics, G., Kimura, M., and Czigler, I. (2011). Visual mismatch negativity reveals automatic detection of sequential regularity violation. *Front. Hum. Neurosci.* 5:46. doi: 10.3389/fnhum.2011.00046

Stothart, G., and Kazanina, N. (2013). Oscillatory characteristics of the visual mismatch negativity: what evoked potentials aren't telling us. *Front. Hum. Neurosci.* 7:426. doi: 10.3389/fnhum.2013.00426

Strain, E., Patterson, K., and Seidenberg, M. S. (1995). Semantic effects in single-word naming. *J. Exp. Psychol. Learn. Mem. Cogn.* 21, 1140–1154. doi: 10.1037/0278-7393.21.5.1140

Tallon-Baudry, C. (2004). Attention and awareness in synchrony. *Trends Cogn. Sci.* 8, 523–525. doi: 10.1016/j.tics.2004.10.008

Tallon-Baudry, C., Bertrand, O., Delpuech, C., and Pernier, J. (1997). Oscillatory γ-band (30–70 Hz) activity induced by a visual search task in humans. *J. Neurosci.* 17, 722–734. doi: 10.1523/JNEUROSCI.17-02-00722.1997

Thut, G., Miniussi, C., and Gross, J. (2012). The functional importance of rhythmic activity in the brain. *Curr. Biol.* 22, R658–R663. doi: 10.1016/j.cub.2012.06.061

Tolentino, L. C., and Tokowicz, N. (2009). Are pumpkins better than heaven? An ERP investigation of order effects in the concrete-word advantage. *Brain Lang.* 110, 12–22. doi: 10.1016/j.bandl.2009.01.001

Tsai, P.-S., Yu, B. H. Y., Lee, C.-Y., Tzeng, O. J. L., Hung, D. L., and Wu, D. H. (2009). An event-related potential study of the concreteness effect between Chinese nouns and verbs. *Brain Res.* 1253, 149–160. doi: 10.1016/j.brainres.2008.10.080

Tugin, S., Hernandez-Pavon, J. C., Ilmoniemi, R. J., and Nikulin, V. V. (2016). Visual deviant stimuli produce mismatch responses in the amplitude dynamics of neuronal oscillations. *Neuroimage* 142, 645–655. doi: 10.1016/j.neuroimage.2016.07.024

Wang, K. (2011). An electrophysiological investigation of the role of orthography in accessing meaning of Chinese single-character words. *Neurosci. Lett.* 487, 297–301. doi: 10.1016/j.neulet.2010.10.041

Wang, L., and Bastiaansen, M. (2014). Oscillatory brain dynamics associated with the automatic processing of emotion in words. *Brain Lang.* 137, 120–129. doi: 10.1016/j.bandl.2014.07.011

Wang, X., Wu, Y., Liu, A., and Wang, P. (2013). Spatio-temporal dynamics of automatic processing of phonological information in visual words. *Sci. Rep.* 3:3485. doi: 10.1038/srep03485

Wei, D., Dowens, M., and Guo, T. (2018). Early lexical processing of written Chinese as indexed by visual mismatch negativity effects. *Sci. Rep.* 8:1289. doi: 10.1038/s41598-018-19394-y

West, W. C., and Holcomb, P. J. (2000). Imaginal, semantic, and surface-level processing of concrete and abstract words: an electrophysiological investigation. *J. Cogn. Neurosci.* 12, 1024–1037. doi: 10.1162/08989290051137558

Wirth, M., Horn, H., Koenig, T., Razafimandimby, A., Stein, M., Mueller, T., et al. (2008). The early context effect reflects activity in the temporo-prefrontal semantic system: evidence from electrical neuroimaging of abstract and concrete word reading. *Neuroimage* 42, 423–436. doi: 10.1016/j.neuroimage.2008.03.045

Xiao, X., Zhao, D., Zhang, Q., and Guo, C.-Y. (2012). Retrieval of concrete words involves more contextual information than abstract words: multiple components for the concreteness effect. *Brain Lang.* 120, 251–258. doi: 10.1016/j.bandl.2011.09.006

Yordanova, J., Falkenstein, M., Hohnsbein, J., and Kolev, V. (2004). Parallel systems of error processing in the brain. *Neuroimage* 22, 590–602. doi: 10.1016/j.neuroimage.2004.01.040

Zhang, J. X., Xiao, Z., and Weng, X. (2012). Neural evidence for direct meaning access from orthography in Chinese word reading. *Int. J. Psychophysiol.* 84, 240–245. doi: 10.1016/j.ijpsycho.2012.02.013

Zhang, Q., Guo, C.-Y., Ding, J.-H., and Wang, Z.-Y. (2006). Concreteness effects in the processing of Chinese words. *Brain Lang.* 96, 59–68. doi: 10.1016/j.bandl.2005.04.004

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 18 March 2019
doi: 10.3389/fpsyg.2019.00513

Text-Based Detection of the Risk of Depression

Jana M. Havigerová[1]*, Jiří Haviger[2]*, Dalibor Kučera[3] and Petra Hoffmannová[1]

[1] Institute of Psychology, Masaryk University, Brno, Czechia, [2] Department of Informatics and Quantitative Methods, University of Hradec Králové, Hradec Králové, Czechia, [3] Department of Pedagogy and Psychology, University of South Bohemia, České Budějovice, Czechia

This study examines the relationship between language use and psychological characteristics of the communicator. The aim of the study was to find models predicting the depressivity of the writer based on the computational linguistic markers of his/her written text. Respondents' linguistic fingerprints were traced in four texts of different genres. Depressivity was measured using the Depression, Anxiety and Stress Scale (DASS-21). The research sample (N = 172, 83 men, 89 women) was created by quota sampling an adult Czech population. Morphological variables of the texts showing differences (M-W test) between the non-depressive and depressive groups were incorporated into predictive models. Results: Across all participants, the data best fit predictive models of depressivity using morphological characteristics from the informal text "letter from holidays" (Nagelkerke r^2 = 0.526 for men and 0.670 for women). For men, models for the formal texts "cover letter" and "complaint" showed moderate fit with the data (r^2 = 0.479 and 0.435). The constructed models show weak to substantial recall (0.235 – 0.800) and moderate to substantial precision (0.571 – 0.889). Morphological variables appearing in the final models vary. There are no key morphological characteristics suitable for all models or for all genres. The resulting models' properties demonstrate that they should be suitable for screening individuals at risk of depression and the most suitable genre is informal text ("letter from holidays").

Keywords: depression, genre, morphology, quantitative linguistics, predictive model

OPEN ACCESS

Edited by:
Eva Kehayia,
McGill University, Canada

Reviewed by:
Roel M. Willems,
Radboud University, Netherlands
Jovana Bjekić,
University of Belgrade, Serbia

****Correspondence:***
Jana M. Havigerová
jamary@mail.muni.cz
Jiří Haviger
jiri.haviger@uhk.cz

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *15 April 2018*
Accepted: *21 February 2019*
Published: *18 March 2019*

Citation:
Havigerová JM, Haviger J, Kučera D and Hoffmannová P (2019) Text-Based Detection of the Risk of Depression. Front. Psychol. 10:513. doi: 10.3389/fpsyg.2019.00513

INTRODUCTION

Depression

The 10th Revision of International Classification of Diseases ICD-10, which is the basis for diagnosing mental disorders in the Czech Republic, classifies depression as an affective disorder (mood disorder). The disorder can have three forms: mild, moderate and severe forms of depression. One of the first symptoms is a change in mood toward the negative pole: the individual feels sad, needless, and/or unimportant. The disorder significantly affects self-confidence, which is often reflected in social relationships. It is typically accompanied by vegetative symptoms which can manifest as gastrointestinal problems (nausea, diarrhea etc.), tremors, sweating, or dry mouth. Sleep is also affected: the individual may have problems falling asleep, waking up, or staying asleep through the night. The effects of depression extend beyond the individual patient, with negative impact on patients' employers, spouses, and children (Scott, 1995; Stewart et al., 2003; Sills et al., 2007).

According to the World Health Organization, depression is the most common mental disorder. Currently, 300 million people suffer from depression (WHO, 2017). The prevalence of depression in the adult population (i.e., clinical depression and definite depression) has been reported at approximately 5% across cultures (Molarius and Janson, 2002; Probst et al., 2006; Munce and Stewart, 2007; Romans et al., 2007; Karger, 2014; Klimusová et al., 2016), and approximately 20% in its milder form (partial symptoms, mild depression, and probable depression; Cho et al., 1998; Romans et al., 2007). The same studies also show that depression is almost twice as prevalent among women compared to men (some studies do not confirm these differences, e.g., Angst et al., 2002). The adult group most at risk is the middle-aged population (Murphy et al., 2000; Stordal et al., 2003; Klimusová et al., 2016). Some studies have reported slight differences in prevalence depending on race (e.g., the white population tended to have higher lifetime prevalence than the black population, Somervell et al., 1989) or according to residence (e.g., the prevalence of depression was significantly higher among rural than urban populations at 6.1 versus 5.2%; Probst et al., 2006).

Worldwide, the prevalence of depression in the population is growing, with an increase of 18% between 2005 and 2015 (WHO, 2017). At the same time, early professional intervention results in improvement of psychological symptoms (e.g., lack of self-confidence, rumination, and anticipation of failure) and to elimination of somatic problems (e.g., gastrointestinal problems and sleeping disorders) in 80% of cases (Siu, 2016). Apart from the relief for the individual suffering from depression, intervention and treatment can significantly improve personal and public health as early detection leads to higher chances of returning to personal, social, and economic life (Reavley et al., 2013).

Goals

Early detection of an individual at risk of depression in initial stages and in mild form is beneficial both for the individual and society. Our study contributes to screening of individuals at risk of this disease. The study provides the original way of depression screening based on an analysis of how the writer uses the language to enable the automatic detection of writer's risk of depression.

Related Works

A frequent type of study focusing on the relationship between text variables and mood disorders (e.g., depression) is a case-study. Case studies analyze texts written spontaneously by authors (individuals) suffering from depression. An important example of this approach is Demjén's (2014) analysis of diaries and works of the American writer Sylvia Plath, who was suffering from lifetime depression and committed suicide in 1963 at the age of 30. In her study, Demjén focused on an analysis of metaphors used by people suffering from depression (i.e., metaphors of separation or loss of control). She found that Sylvia Plath used the second person singular when writing about experiences of conflict or separation. A quantitative analysis of whole texts (not only of the area of metaphors) showed that writers suffering from depression tended to use negative words and expressions with quantifiers with extreme poles (e.g., "everything," "nothing," "always," "never"; Demjén, 2014). Similar results were found in an analysis of texts of the traveler and surveyor Henry Hellyer, who committed suicide at the age of 42. This analysis showed that the pronoun used most frequently was in the first person singular, while the use of the first person plural was much scarcer. Hellyer also tended to use negative words more often, like Sylvia Plath (Baddeley et al., 2011). Pennebaker and Chung (2007) analyzed spontaneous texts of two prominent representatives of Al-Qaeda (Zawahiri and Bin Laden), showing a surprising shift in the use of pronouns closely related to social status, individual and group identity, insecurity, and depression changes.

Automatisation of the processing of linguistic data has recently enabled the use of extensive research strategies. For example, Rude et al. (2004) asked 124 female students attending psychology seminars to write an essay about their deepest thoughts and feelings about college. The students also completed the Beck Depression Inventory, according to which they were divided into groups of currently depressed, formerly depressed and never-depressed people. The authors discovered a positive correlation between degrees of depression and use of the word "I" (i.e., pronouns in the first person singular) and a significantly scarcer use of the pronoun in the second and third persons. It is interesting that other pronouns in the first person singular ("me," "my" and "mine") do not show this correlation. Study based on LIWC analyzes conducted by Lieberman and Goldstein (2006) found that women with breast cancer who used more anger words improved in their health and quality of life, whereas women who used more anxiety words experienced increased depression. Ramírez-Esparza et al. (2008) compared the linguistic markers used by people who write about their depression in internet depression forums with linguistic markers used by people with breast cancer on bbc forums in English and in Spanish. It was found that online depressed writers used significantly more 1st person singular pronouns, less first person plural pronouns in both the English and Spanish forums. Women from depressed forums used less positive emotion words and more negative emotion words than women from breast cancer forums in English and Spanish. Sonnenschein et al. (2018) in their LIWC study provide evidence that the texts of people with mood disorders contain increasingly first-person singular pronouns, depressed as well as anxious, but differ in semantic terms (depressed patients used more words related to sadness). Van der Zanden et al. (2014) found that depression improvement during web-based psychological treatments based on textual communication was predicted by increasing use of 'discrepancy words' during treatment (e.g., would, should – a conditional in Czech language). Self-referencing verbal behavior appears to have specific interpersonal implications beyond general interpersonal distress and depressive symptoms (Zimmerman et al., 2013). A meta-analysis (k = 21, N = 3758) of correlations between first person singular pronoun use and individual differences in depression (which occurs in a number of studies dealing with our topic) were conducted by Edwards and Holtzman (2017) who proven evidence that depression is linked to the use of first person singular pronouns (r = 0.13), this effect is not moderated by demographic factors, such as gender and there is little to no evidence of publication bias in this literature.

Several studies (e.g., Mairesse et al., 2007; Litvinova et al., 2016b) show that indexes combining several studied markers are also important. For example, a reliable predictor of self-destructive behavior (depression is one of the characteristics of such behavior) is the pronominalisation index: the ratio of pronouns to nouns (Litvinova et al., 2016b).

Existing studies do not only show relationships between the way of writing a text and mood disorders (e.g., depression and associated symptoms), but also a reciprocal healing effect of writing certain types of texts. For example, Sayer et al. (2015) conducted an experiment to test the effects of different writing styles using a sample of 1,292 Afghanistan and Iraq war veterans with self-reported reintegration difficulty. In their experiment, veterans who were instructed to write expressively experienced greater reductions in physical complaints, anger, and distress compared to veterans who were instructed to write factually, and, moreover, both writing groups showed reductions in PTSD symptoms and reintegration difficulty compared to veterans who did not write at all. The correlation between occurrence of words and successful intervention was also documented by Alvarez-Conrad et al. (2001).

Studies in clinical psychology clearly show that research on the relationship between the user of a language (e.g., speaker or writer) and their text is meaningful and has potential for the future. A worldwide and rapidly developing approach is the detection of the personality of authors from their texts, involving the design of predictive models based on correlations between quantifiable text parameters and individual psychological traits (Mairesse et al., 2007; Litvinova et al., 2016b). The present study was designed to add to this body of research.

THE AIM OF THE PRESENT RESEARCH

The study presented here focuses on discovery of the relationship between linguistic characteristics of a written text and the level of the emotional state of depression (depressivity) of its author. The focus is on non-content (non-semantic) computational linguistic markers of a written text. The main objective of the study is to find out which texts (and whether or not) can predict depression and what linguistic characteristics are involved in the eventual model. The key step is to create and evaluate predictive models to detect individuals at risk of depression from written texts. Into the models it is necessary to insert only a limited number of variables (DeVaus, 2002), therefore we carried out a two-stage reduction of input linguistic characteristics: (1) the variables having a low variability will be excluded, (2) the variables that can not distinguish between the depressive and non-depressive respondents will be excluded.

Gender differences: Due to the fact that there are gender differences in depression (Murphy et al., 2000; Stordal et al., 2003; Herring and Paolillo, 2006; Johannsen et al., 2015; Klimusová et al., 2016; Rafi, 2019) as well as gender differences in text processing (Litvinova et al., 2017) our analyses are conducted separately for men and women. We expect the results in each of the samples to differ in some features based on the gender of the writer.

Genre differences: Quantitative linguistic markers of a text are affected by the genre (Douglas, 1992; Stamatatos et al., 2000; Herring and Paolillo, 2006). Thus, the analyses are conducted on texts of four different genres. The genres are divided into two categories: formal (cover letter TXT1 and complaint TXT3) and informal texts (letter from holidays TXT2 and letter of apology TXT4).

Necessity and innovativeness of conducting the present study in Czech: Most research on the relationship between the linguistic properties of a text and its author's personal traits have been conducted with texts in English (e.g., Pennebaker's studies above). There is also research on texts in Chinese, Arabic, Spanish, Dutch, French, German, Italian, Russian, Turkish, and Serbian (e.g., Bjekić et al., 2014; Sikos et al., 2014; Sboev et al., 2016). According to Parkvall (2007), Czech is spoken by relatively few native speakers. With 10 million native speakers, Czech is the 83rd most used language in the world and the 15th most frequently used language on the internet. In the studied context, Czech is an under-researched language (W3Techs, 2017) and, with the exception of our own preliminary research, we are aware of no other published research on the relationship between linguistic markers of a text and its writer's personality in Czech.

MATERIALS AND METHODS

Measures

Depression, Anxiety and Stress Scale - 21 Items (DASS-21)

The DASS-21 is a set of three self-report scales designed to measure the emotional states of depression, anxiety and stress. Each of the three subscales contains seven items. Each item is scored on a 4-point scale (0 = *did not apply to me at all*; 3 = *applied to me very much or most of the time*). Thus, a respondent can get 0 to 21 points for each subscale. The DASS-21 is based on a dimensional rather than a categorical conception of psychological disorders. The assumption on which the scale was developed (and which was confirmed by the research data) is that the differences between depression, anxiety and stress experienced by normal subjects and clinical populations are essentially differences of degree (Lovibond and Lovibond, 1995). In this study, we work with the subscale of depression, which assesses dysphoria, hopelessness, devaluation of life, self-deprecation, and lack of interest/involvement, anhedonia, and inertia. In our study, we work with either the total score (0–21 points achieved) or with the cut-off score (non-depressive ≤ 6, depressive > 6), see Lovibond and Lovibond (1995).

Four fictive letters were written on a computer in a pre-defined electronic interface. All four letters were written by each participant. The recommended length of texts was 180–200 words. The participants could see the number of words used on the monitor. However, length was recommended, not strictly prescribed. The content of the text could be entirely fictional.

The sequence of the four scenarios was selected randomly and each scenario was described to participants as follows:

Cover letter (TXT1, formal, positive sentiment): "You have found a job offer that captivated your interest and you aspire to be hired for this position. Therefore, you are going to write a letter to the company's director as a response to his/her offer trying to persuade the director that you are the right candidate for this position."

Letter from holidays (TXT2, informal, positive sentiment): "You are enjoying your time on an amazing vacation. Everything is going well, as expected, and you fully indulge in your popular activities. Therefore, you have decided to write a letter to your friend and convince him/her to come over and enjoy this perfect time with you."

Complaint (TXT3, formal, negative sentiment): "Until recently, you were satisfied with living in your apartment (or your house), not missing anything. Nevertheless, recent issues have made a hell out of a pleasant living. Although you originally strived to sort out the issues in a polite way, it did not help. Therefore, you decided to write an official complaint to the appropriate authorities."

Letter of apology (TXT4, informal, negative sentiment): "You have done something that substantially harmed your relationship with a person you were very close to for a long time. You had promised something that you did not fulfill. You feel sorry and you know that you made a mistake. Because you do not want to lose your close friend, you have decided to write a letter of apology to him/her."

The analyses were conducted on 688 texts that create a corpus of 99,481 words. In all texts, quantitative linguistic variables on various levels of classification (e.g., number of all adjectives, number of superlative forms of adjectives, number of words in singular, etc.) were automatically detected in the process of lemmatization with morphological tagging (Jelínek and Petkevič, 2011).

Quantitative linguistic variables are included in the analyses in the form of relativized isolated features (ratios) and compound indicators (special metrics) as described in the following lists.

Ratios (input = 16 items):

– words per sentence: the number of words divided by the number of sentences,
– lemmas per sqrt words: the number of different lemmas (basic forms) divided by the square root of the number of words,
– sentence complexity: the number of finite verbs divided by the number of sentences,
– punctuations per sentence: the number of punctuation marks divided by the number of sentences,
– exclamation per sentence: the number of exclamation marks divided by the number of sentences,
– AN per ANNA: the number of adjective-noun pairs divided by the number of all pairs (adjective-noun plus noun-adjective),
– colloquial words per sentence: the number of colloquial words divided by the number of sentences,
– singularity index: the number of words in singular divided by the number of all words which have the grammatical category of interest (i.e., divided by the number of singular plus the number of plural plus the number of dual nouns),
– singularity P index: the number of possessive singular words divided by the number of all possessive words,
– vocative index: the ratio of words in vocative to the sum of all other words that have the grammatical category of interest,
– negativity index: the ratio of negative sentences to negative plus affirmative sentences,
– passive index: the number of words in passive divided by the number of words in passive and active,
– imperfectum index: the ratio of perfectum to perfectum plus imperfectum,
– dem per words: the number of diminutive words divided by the number of all words,
– vul per words: the number of vulgarisms divided by the number of all words,
– clq per words: the number of colloquial words divided by the number of all words.

Special metrics (input = 8 items):

– coherence index: calculated using the formula Coh = (particles + conjunctions + prepositions)/(3 * sentence) (Litvinova et al., 2016b),
– pronominalisation index: the ratio of the total number of pronouns to the total number of nouns (Litvinova et al., 2016b),
– formality metric: is calculated using the formula F = (noun + adjective + preposition + article - pronoun - verb – adverb - interjection + 100)/2 (Mairesse et al., 2007),
– trager index: number of verbs/number of adjectives (Sboev et al., 2016),
– readiness to action: number of verbs/number of nouns (Sboev et al., 2016),
– aggressiveness index: number of verbs/number of all words (Sboev et al., 2016),
– activity index: number of verbs / (number of verb + adjective + adverbs),
– autosemantic index: number of autosemantic words (noun, adjective, pronoun, numeral, verb, and adverb) in relation to number of words (Čech et al., 2014).

Data Collection Procedure

Participants were recruited using leaflets and advertisements on social networks. The participants were couples of people older than 15 who enrolled in the study voluntarily. After the study, they were awarded about 50 USD. Data collection was conducted in the controlled environment of a university on weekends from September 2016 to April 2017.

A battery of self-report psychological tests was administered with 4 fictive letters placed randomly between test blocks. The conditions of administration were always identical (the same environment, the same assistants) and relatively naturalistic to make the participants feel comfortable (they were allowed to relax when needed and an assistant was present). The maximum level of structure and identity of situation were strictly obeyed to eliminate the impact of structure of the situation on the correlation between linguistic markers of a text and its writer's personality (e.g., as discussed by Hirsh and Peterson, 2009).

Participants

Quota selection was used to sample participants. The decisive criterion for determination of quotas was age, gender, and education (Škrabal, 2014). The inclusion criteria were Czech citizenship, command of Czech as mother tongue, good psychical condition (without medication with psychopharmaceuticals), good knowledge of each other in each enrolled couple (that would allow the participants to describe each other sincerely and with a detached view). The participants declared fulfillment of conditions by signing a detailed informed agreement.

The sample is made of $N_{\text{resp}} = 172$ respondents, out of whom $n_{\text{m}} = 83$ men, $n_{\text{w}} = 89$ women. The distribution with respect to age and education is given in **Table 1**. The studied properties of the research sample correspond to distribution of the legally competent population in the Czech Republic, which makes generalization of results at this level possible.

Data Analysis

The analysis was conducted in seven steps. *Outlier filtering*: All texts showing outlying values for number of words (<100) and number of sentences (<5) were discarded. *Reduction of variables, step one: low variability variables exclusion*. Descriptive statistics for the studied QL variables were calculated and variables showing low level of variability were discarded from further calculations (if at least one of the following conditions were fulfilled: mdn = 0 or coefficient of variation sd/m < 0.05 or iqr/mdn < 0.05). *Assessment of normality*: A Kolmogorov–Smirnov (K–S) test was conducted. *Lowering granularity of depressive scale*: The depression subscale of the DASS-21 (min–max = 0–21) was transformed to DASS21_01 (non-depressive ≤ 6, depressive > 6), see Lovibond and Lovibond (1995).

Reduction of variables, step two: exclusion variables with non-significant intergroup differences (depressive vs. non-depressive). This comparison was done using the Mann–Whitney test. It was conducted the 8 tests for four genres and two genders. Variables that didn't show significant differences in any of the 8 tests were excluded. *Creation of models*. Eight models were created using logistic regression (for four genres and two genders). Predictive models included only variables showing significant differences between the depressive and non-depressive group in the M–W test (see step 5).

Evaluation of models: The criterion of the quality of each regression model was defined by the Nagelkerke coefficient $r^2 > 0.4$. To assess the predictive power of a model, the following coefficients were calculated: precision (the level of accuracy: the probability that a respondent marked by the model as depressive really is depressive) and recall (also known as sensitivity: the probability that a respondent who really is depressive is classified by the model as depressive). These coefficients are suitable for an unbalanced (unequal) sample (Chawla, 2005). The predictive power of the model was evaluated as sufficient if it met the following conditions: precision >0.8 and recall >0.6. The objectives of the study are exploratory and therefore it is not necessary to use multiple comparison correction (Li et al., 2016). The variables that were inserted into the models went through a two-step selection. Models were further verified according to predefined criteria (precision, recall).

TABLE 1 | Age group, education, and gender of participants ($N_{\text{resp}} = 172$).

			% of Total			Total %
			Education			
Gender			Primary %	High school %	University %	
Male	Age group	Younger than 25	7.2	7.2	1.2	15.7
		25–34	1.2	14.5	6.0	21.7
		35–55	3.6	30.1	8.4	42.2
		Older than 55	2.4	12.0	6.0	20.5
	Total		14.5	63.9	21.7	100.0
Female	Age group	Younger than 25	5.6	5.6	1.1	12.4
		25 to 34	1.1	11.2	5.6	18.0
		35 to 55	2.2	24.7	6.7	33.7
		Older than 55	2.2	27.0	6.7	36.0
	Total		11.2	68.5	20.2	100.0
Total	Age group	Younger than 25	6.4	6.4	1.2	14.0
		25 to 34	1.2	12.8	5.8	19.8
		35 to 55	2.9	27.3	7.6	37.8
		Older than 55	2.3	19.8	6.4	28.5
	Total		12.8	66.3	20.9	100.0

RESULTS

Verification of Assumptions and Preparation of Data for Testing and Creating Models

Outlier Filtering

In each type of text, 1–4 outliers were detected (see **Table 2**), a total of 32 texts did not meet the conditions for entering analysis and were automatically discarded from further calculations. Thus, the analyses consisted of 656 texts.

Reduction of Variables

The stated conditions for variability were fulfilled by 13 quantitative linguistic variables out of the 24 followed ones. These 13 were included in further calculations.

Assessment of Normality

As expected, the studied variables were not normally distributed (K–S, sig < 0.05). Due to the non-normal distribution of data, only non-parametric/non-linear procedures were used in further calculations.

Lowering Granularity of Depressive Scale

The distribution of scores from the depression subscale of the DASS-21 was the following: men (m_m = 5.14, mdn_m = 5, min_m = 0, max_m = 17), women (m_w = 4.04, mdn_w = 3, min_w = 0, max_w = 16). Based on their results on the depression subscale, respondents were divided binarily into non-depressive and depressive groups (see **Table 3**). The cut-off point for these groups (6 points) was derived from the psychometric properties of the test (see Lovibond and Lovibond, 1995). The higher representation of men in the depressive category (about 1/3 of men) compared to women (about 1/5 of women) reflects the characteristics of the research sample (it is unexpected and deserves a separate analysis; let us recall here that the respondents enrolled into the research voluntarily and were included by quota selection. Thus, an identical distribution between men and women and higher prevalence in women were expected.).

TABLE 2 | Frequency of outliers (N_{resp} = 172, N_{text} = 688).

	Male (n_m = 83)		Female (n_f = 89)	
	Non-outliers	**Outliers**	**Non-outliers**	**Outliers**
TXT1	81	2	88	1
TXT2	81	2	85	4
TXT3	81	2	86	3
TXT4	82	1	88	1
Total	*325*	*7*	*347*	*9*

TABLE 3 | Frequency of depression in sample (N_{resp} = 172).

DASS_D	Male (n_m = 83)	Female (n_f = 89)
Non-depressive (D ≤ 6)	65.5%	81.3%
Depressive (D > 6)	34.5%	18.7%
Total	*100*	*100*

Testing of Intergroup Differences (Depressive vs. Non-depressive)

The depressive and non-depressive groups were compared using a Mann–Whitney U test (M–W test, a non-parametric test for independent samples). The test was conducted separately for men and women as well as for each type of text. **Table 4** shows an overview of significances of individual tests (*U* values and mean rank are available from the authors).

There was a significant difference in mean rank between the depressive and non-depressive groups for each of the selected linguistic variables (except for the coherence index) in at least one text. The only exception is the coherence index, which according to the result of M–W test does not differentiate between depressive and non-depressive groups in any of the texts, and thus will be excluded from further calculations. Contrary to our expectations, a higher number of significant differences were found among men compared to women, and more often in formal texts (TXT1 and TXT3).

Creating and Evaluating Models

Eight regression models were created (for four texts among men and four texts among women). **Table 5** presents an overview of significance for individual predictors for each model, with each column representing one model.

The quality of individual models is described in **Table 6** (bold values indicate those model values that meet a predefined quality criterion and allow the model to be accepted).

The only model that fulfilled all the defined criteria is the model created on TXT2 (letter from holidays, informal text with positive sentiment) among women. The stated criteria are approximated by three models for men, namely models based on TXT1 (cover letter), 2 (letter from holiday) and 3 (letter of complaint).

DISCUSSION

The present study focuses on the relationships between linguistic properties of a written text and the level of its writer's currently experienced depressivity (based on the number of points achieved in the DASS-21 test, participants were divided into depressive and non-depressive group, and these two groups were compared). The chosen methodology is novel: (a) the source for analyses were texts written on an assigned topic under strictly controlled experimental conditions (i.e., not spontaneously written texts), (b) only formal, quantitative linguistic syntactical and morphological variables were subject to analyses (not semantic variables, i.e., only the verbal production, not its content were considered), (c) the research sample was representative of an adult population with respect to age and education (quota selection). We have not come across a study conducted using the same methodological basis. Methodologically similar studies are very scarce (e.g., Litvinova et al., 2016b tried to assess the

TABLE 4 | Significance of intergroup differences in ql variables: Mann–Whitney (N_{resp} = 172, N_{text} = 688).

	Male (n_m = 83)				Female (n_f = 89)			
	TXT1	TXT2	TXT3	TXT4	TXT1	TXT2	TXT3	TXT4
Singularity index	0.213	0.175	0.503	0.374	0.120	0.653	0.008	0.459
Singularity_P index	0.039	0.208	0.157	0.470	0.211	0.540	0.542	0.291
Negativity index	0.026	0.000	0.013	0.122	0.664	0.613	0.863	0.399
Words per sentence	0.125	0.266	0.062	0.020	0.676	0.475	0.916	0.750
Sentence complexity	0.000	0.030	0.005	0.004	0.518	0.488	0.578	0.721
Punctuations per sentence	0.130	0.826	0.458	0.612	0.290	0.420	0.487	0.009
Coherence index	0.528	0.613	0.732	0.204	0.747	0.236	0.748	0.530
Pronom index	0.059	0.279	0.006	0.205	0.131	0.711	0.210	0.469
Formality index	0.023	0.086	0.001	0.062	0.112	0.566	0.523	0.317
Trager index	0.022	0.011	0.011	0.032	0.128	0.117	0.846	0.189
Readiness_to_action index	0.014	0.049	0.002	0.058	0.171	0.870	0.781	0.965
Aggressiveness index	0.013	0.063	0.008	0.243	0.234	0.468	0.653	0.910
Activity index	0.011	0.025	0.021	0.259	0.076	0.070	0.474	0.390

TABLE 5 | Predictors of membership in depressive sample: Logistic regression.

	Male (n_m = 83)				Female (n_f = 89)			
	TXT1	TXT2	TXT3	TXT4	TXT1	TXT2	TXT3	TXT4
Singularity index	0.701	0.062	0.845	0.050	0.058	0.630	0.007	0.349
Singularity_P index	0.044	0.630	0.017	0.585	0.112	0.392	0.746	0.728
Negativity index	0.635	0.534	0.401	0.153	0.169	0.071	0.295	0.247
Words per sentence	0.046	0.172	0.250	0.248	0.046	0.059	0.411	0.194
Sentence complexity	0.004	0.089	0.575	0.536	0.198	0.050	0.618	0.201
Punctuations per sentence	0.236	0.210	0.841	0.087	0.818	0.041	0.677	0.798
Pronom index	0.661	0.042	0.017	0.720	0.385	0.682	0.740	0.323
Formality index	0.479	0.631	0.025	0.723	0.525	0.605	0.234	0.236
Trager index	0.639	0.759	0.019	0.841	0.599	0.139	0.186	0.182
Readiness_to_action index	0.333	0.034	0.017	0.688	0.052	0.941	0.506	0.083
Aggressiveness index	0.162	0.418	0.009	0.863	0.060	0.217	0.999	0.567
Activity index	0.374	0.352	0.007	0.694	0.059	0.554	0.622	0.353
Constant	0.431	0.170	0.059	0.637	0.639	0.248	0.086	0.914

Each column represents a single model (N_{resp} = 172, N_{text} = 688).

TABLE 6 | Coefficients of model quality and predictive power: Logistic regression (N_{resp} = 172, N_{text} = 688).

	Male (n_m = 83)				Female (n_f = 89)			
	TXT1	TXT2	TXT3	TXT4	TXT1	TXT2	TXT3	TXT4
Nagelkerke r^2	0.479	0.526	0.435	0.310	0.330	0.670	0.324	0.284
Recall	0.607	0.571	0.556	0.500	0.235	0.800	0.267	0.267
Precision	0.739	0.889	0.789	0.647	0.571	0.889	0.667	0.667

probability of self-destructive behavior of an individual via formal parameters of their texts).

One of the difficult questions was the choice of linguistic variables to include in the models. We have decided for a statistics-based procedure. In the first step, we have excluded variables with low variability. Sufficient variability has been proven for 6 of 16 selected single morpho-syntactic variables: the number of words per sentence, number of finite verbs per sentence, number of punctuation marks per sentence, proportional variables of relative occurrence of singular, possessive singular, negativity, and for all 8 indexes consisting of combinations and ratios of more morpho-syntactic characteristics: index of coherence, pronominalisation, formality, trager, readiness to action, aggressiveness and activity. This means that, in our study, only a limited amount of selected single morpho-synaptic characteristics was found to be suitable for use in distinguishing between non-depressive and non-depressive texts because of low variability; while all indexes showed sufficient variability. These results can support opinion that it is suitable to use indexes combining in formulas more morpho-syntactic characteristics of a text rather than focus attention on each of the observed linguistic characteristics as an insular unit when looking for relationships between a text and the characteristics of the writer of the text, as some other researchers stated (e.g., Litvinova et al., 2016b).

In step two it was verified discriminatory power of each from these 13 variables via M–W test. The results confirmed that all proposed variables have sufficient discriminatory power to distinguish between texts (always at least one of the texts) of non-depressive and depressive people, except one. The exception is index of coherence. Contrary to our expectations, the present study does not validate the index of coherence (Litvinova et al., 2016b) as a suitable predictor of depression. This index is calculated as the sum of particles plus conjunctions plus prepositions divided by 3 times multiple of number of sentences. We believe that the reason why the index of coherence does not differentiate between non and depressive sample lies in the fact that the index includes the synsemantic parts of speech only. This explanation mirrors Pennebaker's (2011) argument that personality is most closely related to pronouns and other autosemantic words than synsemantic ones.

Thirteen linguistic variables (6 single morho-syntactic characteristics, 7 indexes combining more morphosyntactic characteristics) were included into the predictive models. Eight predictive models (for 4 different texts and 2 genders) were created and compared with each other. The results show that acceptable level of accuracy show models predicting depression in men sample from texts TXT1 (cover letter), TXT2 (letter from holidays) and TXT3 (complaint), and in women sample from TXT2 (letter from holidays). Across these 4 models, the probability that an individual will be detected as depressive when he/she is not (type II error) is lower than 0.2. The models for men sample show lower quality in criterion recall (their power to detect a depressive individual) than models for women sample. In other words, models built on texts written by men are more likely to fail to detect an individual with depression (type I error) than to erroneously classify an individual as depressive (type II error). Based on these results, it seems justified to state that, pursuant to the morpho-syntactic characteristics of the text, it is more confident to identify depressive women than depressive men.

For explanation, we need to look at gender differences in general and in our study as well. Most current studies show that women experience more depression than men do (e.g., Munce and Stewart, 2007; Klimusová et al., 2016) or the level of depression occurrence is the same for both men and women (Piccinelli and Wilkinson, 2000). However, in our study, men showed a higher level of currently experienced depression than women – it is opposite, unexpected trend. This might be a hit-or-miss feature of our research sample, the unexpected result of self-nomination sampling strategy. Previously diagnosed mental illness has been set as an exclusion criterion for self-nomination into our research no-clinical sample. Because men go to doctors with psychological problems less often than women (e.g., Angst et al., 2002 show that 48% of men and 59% of women with depression seek a doctor), the women with the same intensive depressive symptoms have been visited their doctor and the previous diagnosis made them unable to enter the research as a non-clinical population. It is possible interpretation why there are more depressive men than women in our research sample, even though the prevalence of depression in men is generally lower. However, this circumstance does not explain why predictive model of women sample is stronger than men's sample predictive models.

The literature has repeatedly described that men and women generally differ in the preference of using some linguistic morpho-syntactic elements in their texts (e.g., Koppel et al., 2002; Argamon et al., 2003; Herring and Paolillo, 2006; Newman et al., 2008; Tausczik and Pennebaker, 2010; Rafi, 2019). Litvinova et al. (2016a) found some text parameters as reliable gender predictors: type-token ratio, formality index, a proportion of prepositions and pronoun-like adjectives, proportion of 100 most frequent words and ration of function (synsemnatic) words to content (autosemantic) words, some of them we operate too. Johannsen et al. (2015) presented a large-scale study of syntactic variation across 11 languages and found that there some universal gender-specific variations across languages: men seem to use numerals and nouns more than women, whereas women use pronouns and verbs more often, men use nominal compounds more often than women. From this point of view, the differences between models found in our study are understandable.

In our study, for women the model predicted depression from TXT2 (letter from holidays: informal text with positive situational sentiment) was of apparently higher quality then models predicting from other texts, while predictive models for men were of comparable quality (with respect to recall and precision metrics) across all the different texts. This result could be related to findings in other research. In this context Biber (1988) stated that women tend to express themselves in the form of "involved" writing while men prefer "informative" writing. Argamon et al. (2003) proved that women tend to present things in a relational way, while men in a non-fictional style of writing. Both cited facts may be related to the fact that the strongest predictive model was found in women just in the text, which is informal – closer to the natural way of verbal presentation of women.

Nor does this knowledge help explain differences found; it provides clear that models predicting the depression differ depending on gender, and that in the future it is necessary to take into account the moderative/mediatory influence of the writer/speaker's gender in modeling relationships between depression (or other personality characteristics) and text.

Overall, our results indicate that TXT4 (letter of apology) is not a suitable text for creation of a reliable and accurate model predicting depression. On the contrary, TXT2 (letter from holidays) seems to be suitable for creating a good fit predictive model for both men and women.

Limitations

The present study was conducted on a quota-selected sample of Czech native speakers. Generalization of the findings to Slavonic languages requires further research and generalization to non-Slavonic languages is not recommended. An unexpected limit of this study is the higher percentage of depressive men in the research sample. Due to the relatively small size of the research sample, we did not further verify the results (e.g., by using split half or cross validation).

CONCLUSION

The leading motivation for our research is to find ways to use automatic analysis of texts (such as cover letters, letters from holidays, blogs, and comments on social networks) to create predictive models that will reliably detect individuals at risk of a mental disorder (such as depression in the present study) so that they can be provided with help as early as possible. In the present study, we calculated four regression models to predict a higher emotional state of depression. The quality of our models indicates that depression can be predicted from informal text written about a holiday and that the quantitative linguistic characteristics that are most strongly suited to the proposed models for men are the pronominalisation index (the ratio of pronouns to nouns) and readiness to action index (the ratio of verbs to nouns) and for women are sentence complexity (the ratio of finite verbs to number of sentences) and punctuation (the ratio of number of punctuation to number of sentences). We plan to extend our future research to a clinical population to analyze the texts of people with a diagnosed mental disorder, especially with depression or phobias. Given our results and the results of other research (e.g., Rude et al., 2004) we plan to pay more attention to autosemantic words, especially to various types of pronouns.

DATA AVAILABILITY

The datasets generated and analyzed in this study are available on request from the authors of the article. If interested, please contact the corresponding author.

ETHICS STATEMENT

The project was approved by the Ethics committee of the University of South Bohemia (headed by professor Hana Šantrůčková, president of the Ethics committee of the USB) that confirmed the project was carried out in accordance with the recommendations of the Ethical code of the University of South Bohemia in České Budéjovice. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

AUTHOR CONTRIBUTIONS

DK brought the original idea and was the main solver of the grant project. JMH and JH formulated the goals of the study and all authors studied and discussed the relationship between the text and its writer. DK arranged a complete collection of quota sample data. JMH arranged a complete collection of clinical sample data. DK, JH, and JMH designed the online collection, pre-processing, and retention of the data. JMH and PH wrote the introduction. JH designed the data processing procedure, performed all the mathematical and statistical calculations, and described the results. JMH thought out and wrote the discussion and all authors improved the content and formulations of the manuscript. JMH formatted the text and DK ensured professional proofreading and uploading of the manuscript for the review process.

FUNDING

The research Computational Psycholinguistic Analysis of Czech Text (CPACT) is funded by the Czech Science Foundation (GA ČR) grant no. 1619087S.

REFERENCES

Alvarez-Conrad, J., Zoellner, L. A., and Foa, E. B. (2001). Linguistic predictors of trauma pathology and physical health. *Appl. Cogn. Psychol.* 15, 159–170. doi: 10.1002/acp.839

Angst, J., Gamma, A., Gastpar, M., Lépine, J.-P., Mendlewicz, J., and Tylee, A. (2002). Gender differences in depression. *Eur. Arch. Psychiatry Clin. Neurosci.* 252, 201–209. doi: 10.1007/s00406-002-0381-6

Argamon, S., Koppel, M., Fine, J., and Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text* 23, 321–346. doi: 10.1515/text.2003.014

Baddeley, J. L., Daniel, G. R., and Pennebaker, J. W. (2011). How Henry Hellyer's use of language foretold his suicide. *Crisis* 32, 288–292. doi: 10.1027/0227-5910/a000092

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Bjekić, J., Lazarević, L. B., Živanović, M., and Knežević, G. (2014). Psychometric evaluation of the Serbian dictionary for automatic text analysis-LIWCser. *Psihologija* 47, 5–32.

Čech, R., Popescu, I., and Altmann, G. (2014). *Metody kvantitativní analýzy (nejen) básnických textù*. Olomouc: Univerzita Palackého.

Chawla, N. V. (2005). "Data mining for imbalanced datasets: an overview," in *Data Mining and Knowledge Discovery Handbook*, eds O. Maimon and L. Rokach (Boston, MA: Springer), 853–867. doi: 10.1007/0-387-25465-x_40

Cho, M. J., Nam, J. J., and Suh, G. H. (1998). Prevalence of symptoms of depression in a nationwide sample of Korean adults. *Psychiatry Res.* 81, 341–352. doi: 10.1016/s0165-1781(98)00122-x

Demjén, Z. (2014). Drowning in negativism, self-hate, doubt, madness: linguistic insights into Sylvia Plath's experience of depression. *Commun. Med.* 11, 41–54. doi: 10.1558/cam.v11i1.18478

DeVaus, D. (2002). *Analyzing Social Science Data: 50 Key Problems in Data Analysis*. Los angeles, CA: SAGE Publications.

Douglas, D. (1992). The multi-dimensional approach to linguistic analyses of genre variation: an overview of methodology and findings. *Comput. Hum.* 26, 331–345. doi: 10.1007/bf00136979

Edwards, M., and Holtzman, N. S. (2017). A meta-analysis of correlations between depression and first person singular pronoun use. *J. Res. Pers.* 68, 63–68. doi: 10.1016/j.jrp.2017.02.005

Herring, S. C., and Paolillo, J. C. (2006). Gender and genre variation in weblogs. *J. Sociolinguistics* 10, 439–459. doi: 10.1111/j.1467-9841.2006.00287.x

Hirsh, J. B., and Peterson, J. B. (2009). Personality and language use in self-narratives. *J. Res. Pers.* 43, 524–527. doi: 10.1016/j.jrp.2009.01.006

Jelínek, T., and Petkevič, V. (2011). "Systém jazykového značkování současné psané češtiny," in *Korpusová lingvistika: Gramatika a značkování korpusù*, Vol. 3, ed. F. Čermák (Praha: Nakladatelství Lidové noviny), 154–170.

Johannsen, A., Hovy, D., and Søgaard, A. (2015). "Cross-lingual syntactic variation over age and gender," in *Proceedings of the 19th Conference on Computational Language Learning*, Beijing, 103–112.

Karger, A. (2014). Geschlechtsspezifische aspekte bei depressiven erkrankungen. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* 57, 1092–1098. doi: 10.1007/s00103-014-2019-z

Klimusová, H., Burešová, I., Dosedlová, J., and Jelínek, M. (2016). "Association of optimism and health-related behavior with mental health in Czech adolescents," in *The European Proceedings of Social & Behavioural Sciences*, eds Z. Bekirogullari, M. Y. Minas, and R. X. Thambusamy (Amsterdam: Future Academy), 598–605. doi: 10.15405/epsbs.2016.11.62

Koppel, M., Argamon, S., and Shimoni, A. (2002). Automatically categorizing written texts by author gender. *Literary Linguist. Comput.* 17, 401–412. doi: 10.1093/llc/17.4.401

Li, G., Taljaard, M., Van den Heuvel, E. R., Levinne, M. A. H., Cook, D. J., Wells, G. A., et al. (2016). An introduction to multiplicity issues in clinical trials: the what, why, when and how. *Int. J. Epidemiol.* 46, 746–755. doi: 10.1093/ije/dyw320

Lieberman, M. A., and Goldstein, B. A. (2006). Not all negative emotions are equal: the role for emotional expression in online support groups for women with breast cancer. *Psychooncology* 15, 160–168. doi: 10.1002/pon.932

Litvinova, T., Seredin, P., Litvinova, O., and Zagorovskaya, O. (2017). "Differences in type-token ratio and part-of-speech frequencies in male and female Russian written texts," in *Proceedings of the Workshop on Stylistic Variation*, Copenhagen, 69–73. doi: 10.18653/v1/w17-4909

Litvinova, T., Seredin, P., Litvinova, O., Zagorovskaya, O., Sboev, A., Gudovskih, D., et al. (2016a). "Gender prediction for authors of Russian texts using regression and classification techniques," in *Proceedings of the Third International Workshop on Concept Discovery in Unstructured Data (CDUD 2016)*, Moscow, 44–54.

Litvinova, T., Zagorovskaya, O., Litvinova, O., and Seredin, P. (2016b). Profiling a set of personality traits of a text's author: a corpus-based approach. *Speech Comput.* 9811, 555–562. doi: 10.1007/978-3-319-43958-7_67

Lovibond, S. H., and Lovibond, P. F. (1995). The structure of negative emotional states: comparison of the depression anxiety stress scales (DASS) with the beck depression and anxiety inventories. *Behav. Res. Ther.* 33, 335–343. doi: 10.1016/0005-7967(94)00075-u

Mairesse, F., Walker, M. A., Mehl, M. R., and Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Intell. Res.* 30, 457–500.

Molarius, A., and Janson, S. (2002). Self-rated health, chronic diseases, and symptoms among middle-aged and elderly men and women. *J. Clin. Epidemiol.* 55, 364–370. doi: 10.1016/s0895-4356(01)00491-7

Munce, S. E., and Stewart, D. E. (2007). Gender differences in depression and chronic pain conditions in a national epidemiologic survey. *Psychosomatics* 48, 394–399. doi: 10.1176/appi.psy.48.5.394

Murphy, J. M., Laird, N. M., Monson, R. R., Sobol, A. M., and Leighton, A. H. (2000). Incidence of depression in the Stirling County study: historical and comparative perspectives. *Psychol. Med.* 30, 505–514. doi: 10.1017/s0033291799002044

Newman, M. L., Groom, C. J., Handelman, L. D., and Pennebaker, J. W. (2008). Gender differences in language use: an analysis of 14,000 text samples. *Discourse Process.* 45, 211–236. doi: 10.1080/01638530802073712

Parkvall, M. (2007). *Nationalencyklopedin*. Available at: https://en.wikipedia.org/wiki/Nationalencyklopedin

Pennebaker, J. W. (2011). *The Secret Life of Pronouns: What Our Words Say About Us*. New York, NY: Bloomsbury Press.

Pennebaker, J. W., and Chung, C. K. (2007). "Computerized text analysis of Al-Qaeda transcripts," in *A Content Analysis Reader*, eds K. Krippendorff and M. Bock (Thousand Oaks, CA: Sage), 2007.

Piccinelli, M., and Wilkinson, G. (2000). Gender differences in depression: critical review. *Br. J. Psychiatry* 177, 486–492. doi: 10.1192/bjp.177.6.486

Probst, J. C., Laditka, S. B., Moore, C. G., Harun, N., Powell, M. P., and Baxley, E. G. (2006). Rural-urban differences in depression prevalence: implications for family medicine. *Fam. Med.* 38, 653–660.

Rafi, M. S. (2019). *SMS Text Analysis: Language, Gender and Current Practices*. https://www.tesol-france.org/uploaded_files/files/Coll07-Shaban-Rafi.pdf

Ramírez-Esparza, N., Chung, C. K., Kacewicz, E., and Pennebaker, J. W. (2008). "The psychology of word use in depression forums in English and in Spanish: testing two text analytic approaches," in *Proceedings of the Second International Conference on Weblogs and Social Media, ICWSM 2008*, Seattle, WA.

Reavley, N. J., Morgan, A. J., and Jorm, A. F. (2013). Development of scales to assess mental health literacy relating to recognition of and interventions for depression, anxiety disorders and schizophrenia/psychosis. *Aust. N. Z. J. Psychiatry* 48, 61–69. doi: 10.1177/0004867413491157

Romans, S. E., Tyas, J., Cohen, M. M., and Silverstone, T. (2007). Gender differences in the symptoms of major depressive disorder. *J. Nerv. Ment. Dis.* 195, 905–911. doi: 10.1097/nmd.0b013e3181594cb7

Rude, S., Gortner, E., and Pennebaker, J. W. (2004). Language use of depressed and depression-vulnerable college students. *Cogn. Emot.* 18, 1121–1133. doi: 10.1080/02699930441000030

Sayer, N. A., Noorbaloochi, S., Frazier, P. A., Pennebaker, J. W., Orazem, R. J., Schnurr, P. P., et al. (2015). Randomized controlled trial of online expressive writing to address readjustment difficulties among U.S. Afghanistan and Iraq war veterans. *J. Trauma. Stress* 28, 381–390. doi: 10.1002/jts.22047

Sboev, A., Litvinova, T., Gudovskikh, D., Rybka, R., and Moloshnikov, I. (2016). Machine learning models of text categorization by author gender using topic-independent features. *Procedia Comput. Sci.* 101, 135–142. doi: 10.1016/j.procs.2016.11.017

Scott, J. (1995). Psychotherapy for bipolar disorder. *Br. J. Psychiatry* 167, 581–588. doi: 10.1192/bjp.167.5.581

Sikos, J., David, P., Habash, N., and Faraj, R. (2014). "Authorship analysis of inspire magazine through stylometric and psychological features," in *Proceedings of the 2014 IEEE Joint Intelligence and Security Informatics Conference*, Hague. doi: 10.1109/jisic.2014.15

Sills, M. R., Shetterly, S., Xu, S., Magid, D., and Kempe, A. (2007). Association between parental depression and children's health care use. *Pediatrics* 119, e829–e836.

Siu, A. L. (2016). Screening for depression in children and adolescents: US Preventive Services Task Force recommendation statement. *JAMA* 315, 380–387. doi: 10.1542/peds.2015-4467

Škrabal, J. (2014). *Úroveň Vzdělání Obyvatelstva Podle Výsledkù Sčítání Lidu*. Available at: https://www.czso.cz/documents/10180/20536250/17023214.pdf/7545a15a-8565-458b-b4e3-e8bf43255b12?version=1.1

Somervell, P. D., Leaf, P. J., Weissman, M. M., Blazer, D. G., and Bruce Livingston, M. L. (1989). The prevalence of major depression in Black and White adults in five United States communities. *Am. J. Epidemiol.* 130, 725–735. doi: 10.1093/oxfordjournals.aje.a115394

Sonnenschein, A. R., Hofmann, S. G., Ziegelmayer, T., and Lutz, W. (2018). Linguistic analysis of patients with mood and anxiety disorders during cognitive behavioral therapy. *Cogn. Behav. Ther.* 47, 315–327. doi: 10.1080/16506073.2017.1419505

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Comput. Linguist.* 26, 471–495. doi: 10.1162/089120100750105920

Stewart, W. F., Ricci, J. A., Chee, E., Hahn, S. R., and Morganstein, D. (2003). Cost of lost productive work time among US workers with depression. *JAMA* 289, 3135–3144.

Stordal, E., Mykletun, A., and Dahl, A. A. (2003). The association between age and depression in the general population: a multivariate examination. *Acta Psychiatr. Scand.* 107, 132–141. doi: 10.1034/j.1600-0447.2003.02056.x

Tausczik, Y. R., and Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* 29, 24–54. doi: 10.1177/0261927x09351676

Van der Zanden, R., Curie, K., Van Londen, M., Kramer, J., Steen, G., and Cuijpers, P. (2014). Web-based depression treatment: associations of clients×

word use with adherence and outcome. *J. Affect. Disord.* 160, 10–13. doi: 10.1016/j.jad.2014.01.005
W3Techs (2017). *Usage of Content Languages for Websites*. Available at: https://w3techs.com/technologies/overview/content_language/all
WHO (2017). *Depression*. Available at: http://www.who.int/mediacentre/factsheets/fs369/en
Zimmerman, J., Wolf, M., Bock, A., Peham, D., and Benecke, C. (2013). The way we refer to ourselves reflects how we relate to others: associations between first-person pronoun use and interpersonal problems. *J. Res. Pers.* 47, 218–225. doi: 10.1016/j.jrp.2013.01.008

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Communication

ORIGINAL RESEARCH
published: 19 March 2019
doi: 10.3389/fcomm.2019.00010

The Phonology of Children's Early Words: Trends, Individual Variation, and Parents' Accommodation in Child-Directed Speech

Nina Gram Garmann[1,2]*, *Pernille Hansen*[2]*, *Hanne Gram Simonsen*[2] and *Kristian Emil Kristoffersen*[3]

[1] OsloMet – Oslo Metropolitan University, Oslo, Norway, [2] MultiLing, University of Oslo, Oslo, Norway, [3] Frambu Resource Centre for Rare Disorders, Siggerud, Norway

The mental lexicon is dynamic and changes throughout the lifespan, but how does it begin? Previous research has established that children's first words depend on their communicative needs, but also on their phonetic repertoire and phonological preferences. In this paper, we focus on the phonological characteristics of children's first words, primarily looking at word-initial labials and word length in Norwegian children's first words, as well as at how parents accommodate to child patterns in their speech. Comparing the Norwegian child data with data from children speaking five different languages, we examine how the child's emergent lexicon is on the one hand shaped by the input of the ambient language, but on the other hand limited by more common phonological characteristics of child speech. Based on data from parental reports (CDI), we compared the 50 first words in Norwegian to those in Danish, Swedish, English, and Italian, analyzing two phonological aspects: word initial bilabials and word length in syllables. We found that Norwegian-speaking children follow the children speaking these other languages in having an affinity for word initial bilabials, but that the proportions of mono-, di-, and polysyllables vary depending on the language acquired. Comparisons of the Norwegian child data with samples of adult directed speech (ADS) and child-directed speech (CDS) revealed more word-initial bilabials and shorter words among children than among adults. The CDS was more similar to children's speech than ADS concerning the two phonological aspects dealt with here, which suggests that parents accommodate to children in phonologically detailed ways.

Keywords: phonology, lexicon acquisition, first words, bilabial, word length, Norwegian, communicative development inventories (CDI), CHILDES (Child Language Data Exchange System)

OPEN ACCESS

Edited by:
Victor Kuperman,
McMaster University, Canada

Reviewed by:
Marilyn Vihman,
University of York, United Kingdom
LouAnn Gerken,
University of Arizona, United States

***Correspondence:**
Nina Gram Garmann
nigrag@oslomet.no
Pernille Hansen
pernille.hansen@iln.uio.no

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *31 October 2018*
Accepted: *25 February 2019*
Published: *19 March 2019*

Citation:
Garmann NG, Hansen P, Simonsen HG and Kristoffersen KE (2019) The Phonology of Children's Early Words: Trends, Individual Variation, and Parents' Accommodation in Child-Directed Speech. Front. Commun. 4:10. doi: 10.3389/fcomm.2019.00010

INTRODUCTION

The phonology of children's first words can be influenced by the ambient language on the one hand and by children's cognitive and motoric abilities, such as memory capacity, vision, proprioceptive feedback from the articulators and motoric dispositions and control on the other (Mulford, 1988; de Boysson-Bardies and Vihman, 1991; MacNeilage and Davis, 2000; McCune and Vihman, 2001; Majorano et al., 2014). Disentangling these factors and their influence on children's lexicons is interesting for both practical and theoretical reasons.

Regarding practical matters, any assessment tool that aims for comparability across languages must take cross-linguistic differences into account (Peña, 2007). The need to establish what is language-specific can be illustrated by results from the development of the Cross-linguistic Lexical Tasks (Haman et al., 2015), an assessment tool developed to identify language impairment in multilingual preschoolers. Here, attempts were made to account for phonological complexity across more than 25 different languages through a set of universal criteria. However, there was no stable relationship between this complexity measure and children's performance on the tasks (Haman et al., 2017; Hansen et al., 2017). Hansen et al. (2017) suggested that the measure failed because it did not take cross-linguistic differences into account.

From a usage-based point of view, the relationship between what is specific to children regardless of the ambient language and what is language-specific is also theoretically interesting, as it can shed light on the role of input in acquisition. Given that we build our mental representations of language directly on tokens of language use, properties of the input such as frequency and phonological salience are crucial (Bybee, 2010). However, as young children are limited by cognitive and motoric abilities still under development, it is not given that they are able to take in, and clearly not to reproduce, all the tokens of language that they are exposed to.

In this paper, we will use data collected from a lexical assessment tool that has been adapted to a wide range of different languages: *The MacArthur-Bates Communicative Development Inventories (CDI)*, developed by Fenson et al. (2007). CDI data from large numbers of children have been used to investigate cross-linguistic patterns in children's lexical development (Bleses et al., 2008; Braginsky et al., 2016) as well as the semantics of children's first words (Caselli et al., 1995; Wehberg et al., 2007; Braginsky et al., under review)[1].

The CDI has also been used to analyse phonological acquisition in French and Danish (Gayraud and Kern, 2007; Wehberg et al., 2007). These studies indicate that the first target words Danish children acquire are predominantly monosyllables, whereas French children acquire a balanced proportion of monosyllabic and disyllabic target words. On the other hand, Gayraud and Kern (2007) and Wehberg et al. (2007) report that a large proportion of the early-acquired target words in both Danish and French have a word initial bilabial. Could it be that word length in syllables depends on the ambient language, but that the affinity toward word initial bilabials has more to do with children's universal cognitive, visual, and motoric abilities?

When Wehberg et al. (2007) compared the proportion of word initial bilabials with proportions of other word initial consonants in Danish children's first words, they saw that 45 percent of the words started with an initial bilabial, and that no other initial consonant was as frequent as any of the bilabial consonants. Gayraud and Kern (2007) showed that at 24 months, 45 percent of French children's targeted nouns started with a bilabial. As for word length in syllables, (Wehberg et al., 2007, p. 370) found that only four of the 50 first words in Danish were 'decidedly polysyllabic' in adult pronunciation, i.e., 92 percent may be produced as monosyllables. On the other hand, Gayraud and Kern (2007) showed that at 24 months, French children aimed at 55 percent monosyllables, and 45 percent disyllables.

Wehberg et al. (2007, p. 377) considered word initial bilabials to be universal to child language, but reported that in other respects, children's words are close to their models. This could imply that they think that word length in syllables in children's words corresponds to the ambient language. Gayraud and Kern (2007) looked at the development of children's acquired target nouns over time. According to their findings, early nouns have much in common with babbling, that is, they are typically short, with initial bilabials and open syllables, whereas the nouns become more similar to the ambient language over time, with a diversification of word initial sounds, syllable types, and word length in syllables.

The results from Gayraud and Kern (2007) and Wehberg et al. (2007) are based on analyses of CDI data from 125 to 183 children, respectively. The number of participants is a strength for both studies, but note that parents are only asked to report on which words their children aim at, not on their actual pronunciations. Thus, the CDI findings above need to be supplemented by data covering children's actual productions.

de Boysson-Bardies and Vihman (1991) analyzed consonants in spontaneous speech data in American, French, Swedish and Japanese children aged 9–19 months. Examining the words that the children attempt at, they found that although there were significant differences between the languages, there was a large proportion of word initial labials in all of the four different languages: French, 60%; English, 49%; Swedish, 41%; Japanese, 41% (de Boysson-Bardies and Vihman, 1991, p. 308). Majorano et al. (2014) reported similar results for Italian infants. These results support the idea that word initial bilabials might be a cross-linguistic characteristic of children's early words.

When it comes to word length in syllables, Vihman and Croft (2007 p. 687) reported that in diary and spontaneous speech data, disyllables seem to be the most common word form of early words across languages (Estonian, Finnish, French, Greek, Hebrew, Hindi, Italian, Japanese, Spanish, Swedish, Welsh). In the Germanic languages English, Dutch, and German, however, children aim mostly at monosyllables. The results from Wehberg et al. (2007) indicate that Danish also belongs to the group of Germanic languages where children acquire mostly monosyllabic target words. According to Vihman and Croft (2007), Swedish is an exception to this pattern, because children acquiring Swedish seem to aim at a balanced proportion of mono- and disyllables. In sum, these findings indicate that children have an affinity toward target words with initial bilabial, but that there are cross-linguistic differences in the number of syllables in the targeted words.

Previous research has demonstrated that CDI data are comparable across languages (Caselli et al., 1995; Bleses et al., 2008; Law and Roy, 2008), but this potential has not been fully exploited for phonological purposes. In addition, we do not know how CDI data compare to spontaneous speech data. Across several languages, including Norwegian, children have been found to have individual sound preferences (McCune

[1]Braginsky, M., Yurovsky, D., Marchman, V., and Frank, M. C. (under review). Consistency and variability in word learning across languages. doi: 10.31234/osf.io/cg6ah

and Vihman, 2001; Vihman and Croft, 2007; Garmann and Torkildsen, 2017). We are therefore interested in studying word initial bilabials and word length in syllables in CDI data in more languages as well as comparing CDI data and spontaneous speech data.

The Current Study

In this study, we use CDI data to identify Norwegian children's first words, and compare our findings with published lists of first words based on CDI surveys from English, Italian, Swedish, and Danish (Caselli et al., 1995; Eriksson and Berglund, 1999; Wehberg et al., 2007), to see whether the same phonological tendencies are found in all these languages. The Norwegian CDI data are compared to spontaneous speech data from Norwegian to look at the relationship between aggregated data and individual children's target words as well as actual pronunciations. We expect Norwegian-speaking children to have a large proportion of target words with initial bilabials, but that the proportions of target words and actual pronunciations with initial bilabials may still vary individually.

Do the cross-linguistic differences in word length result from differences in the ambient languages? Vihman et al. (1994) have investigated content words in mothers' speech in English, French and Swedish, reporting predominantly monosyllabic words (69 percent) among English speaking mothers, but a balanced proportion of mono- and disyllabic words in French and Swedish mothers (Vihman et al., 1994, p. 656 Table 4). (Keren-Portnoy et al. (2009), p. 17) state that Italian CDS contains mostly two- or three-syllable words, and suggest that this may be the reason why Italian children target longer words than English children do.

There is no existing research on the proportions of mono- and disyllabic words in Norwegian. As Danish, Swedish, and Norwegian have a common ancestor, Norwegian may be similar to either Danish or Swedish. It is more likely, however, that Norwegian is more similar to Swedish than to Danish when it comes to word length because Danish has undergone severe phonological reductions involving the loss of syllables (Basbøll, 2005, p. 293). As we know that CDS may deviate from ADS (Snow, 1972; Cruttenden, 1994; Englund, 2005; Englund and Behne, 2005), we will look into the phonological characteristics in Norwegian CDS and ADS to highlight relevant differences. Against this background, we will test the following hypotheses:

1. A high proportion of initial bilabials is a property of early words independent of language, and should thus also characterize Norwegian children's first words.
2. The length of early words will vary across languages, and Norwegian children's first words will be balanced between mono- and disyllabic words.
3. Norwegian adults will adapt their speech with respect to both properties in CDS, but still produce fewer words with initial bilabials and longer words than Norwegian infants.

METHODS

To test our hypotheses, we first made a list of the 50 first targeted words in Norwegian based on CDI norms (Kristoffersen and Simonsen, 2012; Kristoffersen et al., 2013; Simonsen et al., 2014).[2] Then we analyzed the proportions of word initial consonants and mono-, di-, and polysyllables in the 50 first targeted words in Norwegian, and investigated the validity of this method by comparing the figures to cross-sectional analyses of the same two characteristics by vocabulary size. We also reanalyzed the lists of first targeted words for Danish, English, Italian, and Swedish (Caselli et al., 1995; Eriksson and Berglund, 1999; Wehberg et al., 2007) to compare the proportions of bilabials and the proportions of monosyllables, disyllables, and polysyllables in all five languages.

The languages that we analyse here represent the two different phonological groups as defined by Vihman and Croft (2007): the general disyllabic group (Italian), and the more Germanic monosyllabic group (Danish and English). We have also included the language that Vihman and Croft (2007) regard as an exception, namely Swedish. It is particularly interesting to compare the three Scandinavian languages to see whether Swedish is an exception to the Germanic pattern as suggested by Vihman and Croft (2007), or whether there is no such Germanic pattern, and rather, that Danish is the odd language out among the Scandinavian languages.

Following the cross-linguistic analyses, we compared Norwegian CDI data with results from Norwegian children's spontaneous speech, using data from video-recorded play sessions between child and parent(s) (Garmann, 2016)[3]. Then, we compared the children's target words as well as actual productions to CDS from the same corpus as well as ADS from one video-recorded conversation between adults in the Norwegian speech corpus *NoTa–Oslo* (University of Oslo, 2013; Hagen and Simonsen, 2014).

Population Based CDI Data

Following the method in Caselli et al. (1995), we pooled data from 2056 children assessed with the Norwegian infant CDI form (*CDI I Words and Gestures*). These data are cross-sectional, and the children's age range from 8 to 20 months, which is a wider range than the one in Caselli et al. (1995) where the range was 8–16 months. Words were considered to be acquired only if they were checked as produced by the child.

The 50 First Words

A list of the 50 first words in Norwegian was extracted by ranking all the words in the CDI checklist by number of occurrences in the CDI responses, and the words were ranked from most common (1) to least common (50). This means that the first words on the list are the words that Norwegian children are most likely to acquire first. To compare the phonological characteristics of the 50 first words in Norwegian with similar words in other languages, we reanalyzed the first words in English, Italian, Danish, and Swedish as listed in Caselli et al. (1995), Eriksson and Berglund (1999), and Wehberg et al. (2007). Whereas the lists of about 50 first words in Norwegian, Danish,

[2]The Norwegian CDI norms are available on wordbank.stanford.edu (Frank et al., 2017).

[3]The corpus can be accessed on doi.org/10.21415/T5P59D.

English, and Italian are the results of analyses of CDI responses from an age range, 8–16 or 8–20 months, the Swedish list in Eriksson and Berglund (1999) consists of the 44 target words that 80 percent or more of all 16-month-olds are reported to produce. We still think the list is comparable to the other lists, since it is based on CDI results and the 16 months stage is included in our Norwegian sample as well as in the other CDI samples.

We have excluded proper names (the child's or a caregiver's name) as well as words used in games or routines only ('peekaboo' and 'patty cake') from the lists in the phonological analyses, because the phonology of these words are unknown to us as they may vary between individuals. We have included onomatopoeia like animal sounds and car sounds—largely these are quite standardized and function as nouns in child (and child directed) speech. Both content words and function words are included. For our phonological analyses, the English list of 50 first words contains 51 words (as two words are ranked as number 50), the Norwegian list contains 49 words, the Italian list 48 words, the Danish list 47 words, and the Swedish list 42 words when names and words used in games and routines are removed.

We have also analyzed the development of the phonological characteristics (the word initial sound and word length in syllables) in Norwegian. The CDI responses with 1–100 words reported as produced were grouped by vocabulary size: 1–10 words, 11–20 words, and so on up to the CDI responses with 91–100 words checked as produced. The proportion of target words with initial bilabials and the proportions of mono-, di-, and polysyllabic target words were calculated for each child, and mean proportions were then calculated within each vocabulary group.

Norwegian Children's Speech

The Garmann (2016) corpus consists of about 60 video recordings from eight monolingual Norwegian-speaking children, four girls and four boys, aged 1;2–2;1, followed longitudinally over a year. Each video recording covers a 30-min session of child–parent interaction at home. The sessions have been transcribed using Phon (Rose et al., 2006).

For each child, one of the parents was asked to fill in a CDI form on the Internet following each recording session. We wanted to compare the phonological characteristics of the first 50 words to spontaneous speech data from children at a similar stage in lexical development. For this reason, we looked for the video recordings that corresponded to the first month in which the parent reported the child to produce at least 50 words in the CDI. For two of the children, this was impossible to determine due to lacking CDI responses. In addition, one of the boys had a sudden jump from 36 words at age 1;8 to 127 words at age 1;9, and we thought it unsuitable to compare data from any of these recordings to the rest of the recordings, where the corresponding CDIs had 54–64 words. For gender balance, we then excluded one of the girls, leaving two girls and two boys for the analyses: Iben (aged 1;6), Johanna (aged 1;3), Marius (aged 1;7), and Olav (aged 1;8). In total, the four children produced 698 word tokens.

Transcription of Child Data

We performed an orthographic transcription of the child data as well as a narrow phonetic transcription with IPA. In total, six transcribers were involved. To check the validity of the orthographic transcriptions, a random 2-min excerpt from each child was transcribed by an independent transcriber, who had access only to the utterance segmentation from the original transcription. The phonetic transcriptions were validated through the same procedure, allowing the second transcriber to see both segmentation and orthographic transcription, but not the phonetic tier.

Transcription agreement was then calculated with respect to the two phonetic properties investigated. As an example, a production of *ballong* 'balloon' transcribed as [naŋɛh] by the original transcriber, but as [wa.ŋə^{h}] by the second transcriber counts as a disagreement about the initial bilabial, but agreement on the number of syllables. In the orthographic transcription, there was a 100% overlap between the transcribers concerning occurrences of initial bilabials, and a 90% agreement on the number of syllables in the utterances transcribed. In the phonetic transcriptions, there was a 99% agreement on initial bilabials (with the one disagreement given as example above), and a 91% agreement on the number of syllables.

Analyses of Child Words

The analyses of target words in children's speech are based on word types for the group as a whole, whereas the analyses of the children's productions are based on word tokens. Both content and function words were included. Most of the children's productions were one-word utterances. Occasionally, the children produced two-word utterances. These were mostly considered as such, but if the prosody indicated that they were produced as unanalyzed units by the child, i.e., there was only one stressed syllable over two words (e.g., Marius: [ɛˈdom] *er tom* 'is empty'), we judged them to constitute single words in the child's vocabulary. In our data set, the group as a whole produced 697 tokens and 101 different target words. The words were analyzed with respect to the word initial sound and the number of syllables in each word, using the same categories as for the CDI data set.

Norwegian CDS and ADS

For analyses of CDS, we used four 30 min' sessions where the parents of Iben, Johanna, Marius, and Olav talked with their children. In total, these four sessions include 6,929 word tokens. Three transcribers in total transcribed the adult utterances orthographically. A fourth transcriber (the first author) transcribed the parents' speech in the first four minutes of each video recording. The transcription agreement, calculated the same way as above, was 92% for the number of syllables and 93% for whether the words had an initial bilabial.

For analyses of ADS, we analyzed the speech of a 28-year-old male from Oslo in conversation with another male. The conversation was taken from the speech corpus *NoTa–Oslo* (University of Oslo, 2013; Hagen and Simonsen, 2014), which consists of interviews of and conversations by speakers of Urban East Norwegian. The corpus is transcribed orthographically.

The conversation lasts for about 40 min, and the 28-year-old's utterances include 4,149 word tokens.

Excerption and Analyses of CDS and ADS Words

There are reasons to believe that children do not give the same attention to all words. Vihman et al. (1994) showed that function words make up maximum 8 percent of the children's words in English, French, and Swedish, and that when it comes to word length and initial consonant, the children's words (including function words) are more similar to content words in CDS than to all of the words in mothers' running speech. They argued that these findings suggest that children attend mostly to content words in the input they receive, and not to unstressed function words. For this reason, we only analyzed the content words in the adult speech, using a wide definition outlined in the **Appendix**. Similarly to Vihman et al. (1994), only CDS from the parents was analyzed, recitations from books were disregarded, and so were imitations of the child's utterances to determine what the child said.

In the CDS recordings, the researcher was involved in the conversations to some degree. For Marius, only the mother was present in addition to the child and the researcher. For Iben, Johanna and Olav, the father was also present. In some of the recordings, visitors or other family members also participated in the conversations. In the analyses of CDS, speech by other speakers, e.g., the researcher or friends, was disregarded. Speech directed toward the researcher (the first author) was also disregarded, as was speech between the parents. If both parents were present, the CDS from both of them was included. We corrected typing errors in the transcriptions, and disregarded fragments. This method resulted in 838 CDS word types from the group as a whole. The ADS recording contained 873 content words types, slightly more than the CDS even though the ADS recording is shorter (40 min vs. 2 h in total for CDS). The phonological analyses were carried out in the same way as for the child data sets.

RESULTS

We initially present the 50 first targeted words in Norwegian, analyse the words with respect to phonological characteristics, and compare the results to reanalyses of the 42–51 first words for English, Italian, Swedish, and Danish in Caselli et al. (1995), Eriksson and Berglund (1999) and Wehberg et al. (2007). Then we take a closer look at Norwegian: First, we investigate the stability of the phonological characteristics by exploring how they vary with vocabulary size. Second, we compare the Norwegian CDI results with targeted words and actual productions in data from spontaneous speech by children. Finally, we look at the role of the ambient speech by comparing children's speech to adults' speech. Exact numbers for the results presented in the figures are published as **Supplementary Material**.

CDI Results

The fifty words that Norwegian children most commonly acquire early are listed in **Table 1**. *Borte!* 'Peek-a-boo!' is only used

TABLE 1 | The 50 first Norwegian target words and the percentage of CDI I responses where the word is checked as produced.

Rank	Word	Phonemic transcription	Translation	% of occ.
1	*mamma*	/^{2}mama/	'mummy'	61.2
2	*hei*	/^{1}hæi/	'hi'	59.2
3	*brrr (bil-lyd)*	/^{1}brrr/	car sound	58.2
4	*pappa*	/^{2}papa/	'daddy'	57.5
5	*nam-nam*	/^{2}namnam/	'yummy'	57.3
6	*nei*	/^{1}næi/	'no'	53.1
7	*ha det*	/^{1}haːde/	'bye-bye'	52.8
8	*bææ*	/^{1}bæː/	sheep sound	50.2
9	*takk*	/^{1}tak/	'thank you'	49.3
10	*voff voff*	/^{2}vovov/	dog sound	45.7
11	*Borte!*	–	peek-a-boo	44.2
12	*ja*	/^{1}jaː/	'yes'	43.7
13	*møø*	/^{1}møː/	cow sound	42.3
14	*au*	/^{1}æʉ/	'ouch'	40.2
15	*ball*	/^{1}bal/	'ball'	39.8
16	*gakk gakk*	/^{2}gakak/	duck sound	35.0
17	*mjau*	/^{1}mjæʉ/	cat sound	34.7
18	*bade*	/^{2}ba:de/	'take a bath'	31.3
19	*lys*	/^{1}ɭy:s/	'light'	30.6
20	*bil*	/^{1}bi:ɭ/	'car'	30.2
21	*banan*	/ba^{1}naːn/	'banana'	29.9
22	*se*	/^{1}seː/	'look'	29.4
23	*baby*	/^{1}beːbi/ or /^{1}bæibi/	'baby'	28.5
24	*hysj*	/^{1}hyʂ/	'shh'	27.3
25	*is*	/^{1}i:s/	'ice cream'	26.7
26	*sko*	/^{1}sku:/	'shoe'	25.6
27	*der*	/^{1}dæ:r/	'there'	25.2
28	*katt*	/^{1}kat/	'cat'	25.0
29	*god natt*	/gu^{1}nat/	'good night'	24.9
29	*bok*	/^{1}bu:k/	'book'	24.9
29	*grr*	/^{1}grrr/	lion sound	24.9
32	*bleie*	/^{2}bɭæie/	'diaper'	24.4
33	*mer*	/^{1}meːr/	'more'	23.6
34	*smokk*	/^{1}smuk/	'pacifier'	22.7
34	*melk*	/^{1}meɭk/	'milk'	22.7
36	*hest*	/^{1}hest/	'horse'	22.5
37	*mat*	/^{1}ma:t/	'food'	22.2
38	*eple*	/^{2}epɭe/	'apple'	22.2
39	*drikke*	/^{2}drike/	'drink'	21.8
40	*den*	/^{1}den/	'that'	21.6
41	*hund*	/^{1}hʉn/	'dog'	21.5
42	*kake*	/^{2}ka:ke/	'cake'	21.3
43	*vann*	/^{1}van/	'water'	20.8
44	*kjeks*	/^{1}çeks/	'cookie'	20.5
45	*nese*	/^{2}ne:se/	'nose'	20.3
46	*borte*	/^{2}buʈe/	'away'	20.2
47	*(leke)bamse*	/^{2}bamse/	'teddy bear'	19.8
48	*øye*	/^{2}øye/	'eye'	19.4
49	*gris*	/^{1}gri:s/	'pig'	18.5
50	*ku*	/^{1}kʉ:/	'cow'	18.2

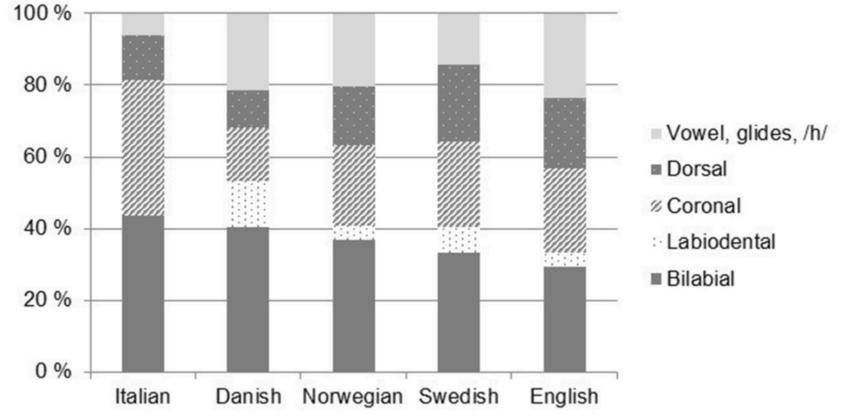


FIGURE 1 | Proportions of place of articulation of the initial sound in the 50 first words in Norwegian compared to Italian, Danish, Swedish, and English based on data from Caselli et al. (1995), Eriksson and Berglund (1999), and Wehberg et al. (2007), sorted by proportion of bilabials.

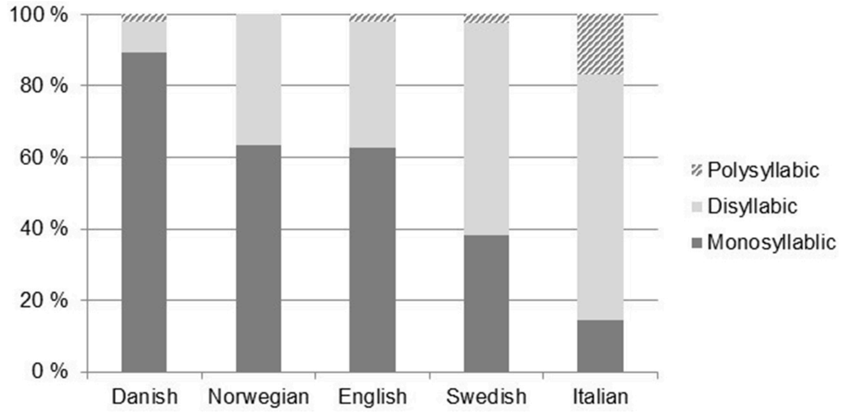


FIGURE 2 | Proportions of mono-, di-, and polysyllabic words among the 50 first target words in Norwegian, compared to percentages in Danish, English, Swedish, and Italian based on data from Caselli et al. (1995), Eriksson and Berglund (1999), and Wehberg et al. (2007), sorted by proportion of monosyllabic words.

in games or routines, and is excluded from the phonological analysis. Among the 49 words left, there are 18 words with initial bilabial, 2 words with initial labio-dental, 11 words with initial coronal, 8 words with initial dorsal, and 10 words with initial vowel, glide or /h/. **Figure 1** shows the proportions of the different places of articulation of word initial consonants in target words in Norwegian compared to Italian, Danish, Swedish, and English. For all the five languages, the largest category of the first words is target words with initial bilabials, ranging between 29 and 44 percent. According to a chi-square test, the differences between the languages are not significant ($\chi^2(4, N = 237) = 2.67, p = 0.61$).

The Norwegian list contains 31 monosyllables, 18 disyllables and no polysyllables. **Figure 2** shows the percentages of monosyllables, disyllables, and polysyllables in Norwegian compared to Danish, English, Swedish, and Italian. The Danish first words list is skewed toward monosyllabic words, whereas the Italian list is skewed toward disyllabic and polysyllabic words. The Norwegian, English, and Swedish lists are balanced between mono- and disyllabic words, although for English and Norwegian, monosyllables are the most frequent, while for Swedish, disyllables are more frequent. A chi-square test shows that there are significant cross-linguistic differences in the proportions of monosyllables in target words ($\chi^2(4, N = 237) = 61.22$, $p < 0.001$); According to pairwise comparisons of the proportions (see **Table 2**), Danish has significantly more monosyllables than the four other languages, and Italian has significantly fewer monosyllables than Danish, English, and Norwegian. There is no significant difference between English, Norwegian, and Swedish, or between Swedish and Italian.

Figures 3, **4** illustrate that for Norwegian, the phonological characteristics in target words are quite stable during the early lexical development. The distribution between monosyllabic and disyllabic words is quite balanced throughout, and the proportion of word-initial bilabials is quite stable in the area between 30 and 40%. Regarding the first 10 words, the items *mamma* 'mommy' and *pappa* 'daddy' contribute to high percentages of disyllables and word-initial bilabials.

TABLE 2 | Levels of significance (p-values) for pairwise comparisons of the proportions of monosyllables in the first 50 target words of each language according to CDI data.

	Danish	English	Norwegian	Swedish
English	0.029	–	–	–
Norwegian	0.028	1.000	–	–
Swedish	<0.001	0.087	0.087	–
Italian	<0.001	<0.001	<0.001	0.084

Comparisons were performed with Holm corrections.

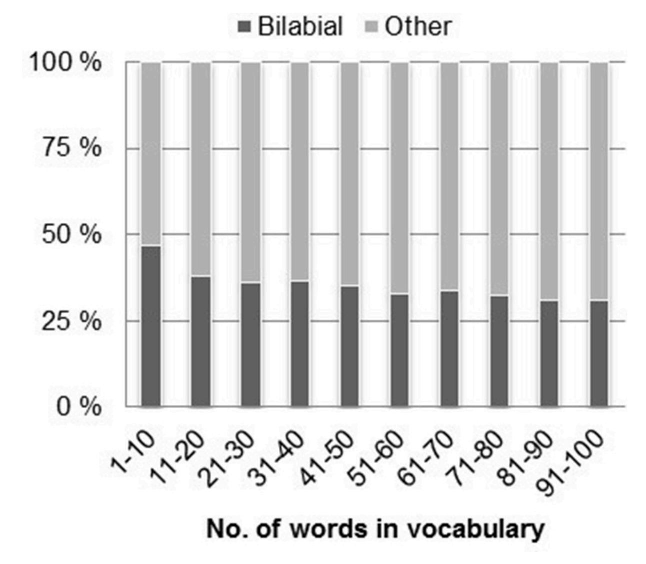


FIGURE 3 | Mean proportions of word initial bilabials by vocabulary size.

Children's Speech Compared to CDS and ADS

In this section, we present the proportions of word initial bilabials and the proportions of mono-, di-, and polysyllables in target words and actual productions in Norwegian children's spontaneous speech, and see whether the proportions are reflected in adults' speech directed to children (CDS) and/or in adults' speech directed to other adults (ADS).

When it comes to place of articulation of word initial sounds, 32 percent of the four Norwegian children's 101 target word types have a word initial bilabial. According to a chi-square test, this is not significantly different from the corresponding proportion in the Norwegian CDI based first words list (37 percent) ($\chi^2(1, N = 150) = 0.18$, $p = 0.67$). In the four children's actual productions, 31 percent of the 697 tokens have a word initial bilabial. According to a chi-square test, this is not significantly different from the 32 percent in the target words ($\chi^2(1, N = 798) = 0.003$, $p = 0.96$). However, there is notable variation between the children: Whereas, Iben produces 51 percent of her word tokens with initial bilabials, Johanna does this in only 8 percent of her word tokens.

Turning to the content words of Norwegian adults' speech, 19 percent of the CDS word types have a word initial bilabial. In ADS, the corresponding proportion is 11 percent. As the proportion of targeted word types with initial bilabials in children's spontaneous speech is almost twice as high as in content words in CDS and almost three times as high as in content words in ADS, this suggests that the high proportion of targeted words with word initial bilabials is in fact typical to children's speech. A chi-square test confirms that the differences between the three data sets are significant ($\chi^2(2, N = 1, 812) = 38.67$, $p < 0.001$). The fact that content words in CDS have a higher proportion of words with initial bilabials than content words in ADS implies that parents adapt to children's preference for these words.

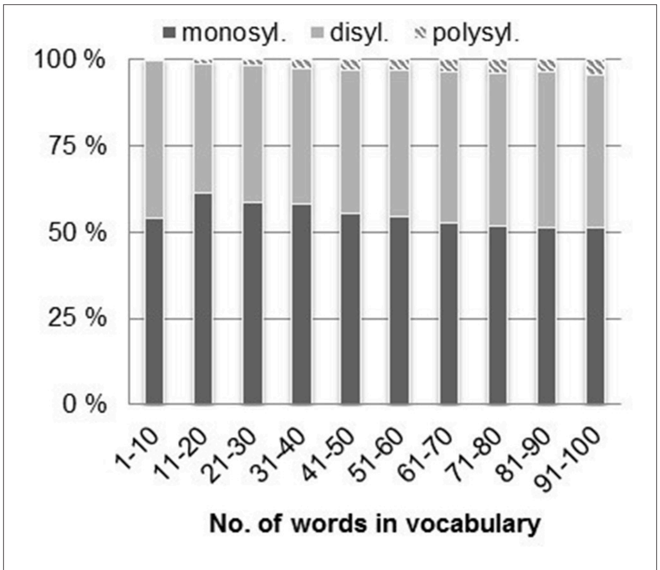


FIGURE 4 | Mean proportions of word length in syllables by vocabulary size.

Concerning word length, our analysis of spontaneous speech data shows that the four Norwegian children in our data set aim at 42 percent monosyllables, 54 percent disyllables, and 4 percent polysyllables in spontaneous speech. Thus, there are fewer monosyllables and more di- and polysyllables in spontaneous speech than in the targeted words in the CDI based first words list (63 percent monosyllables, 37 percent disyllables, and no polysyllables). According to a chi-square test, the difference is significant ($\chi^2(1, N = 140) = 14.71$, $p < 0.001$). In the four children's actual productions, there are 50 percent monosyllables, 45 percent disyllables, and 5 percent polysyllables. According to a chi-square test, this is not significantly different from the proportions in the analysis of the target words [$\chi^2(1, N = 798) = 2.14$, $p = 0.14$]. There is some variation between the children: Johanna produces 33 percent of her word tokens with monosyllables, whereas Olav produces 65 percent monosyllables.

Our analyses of content words in adult speech show that Norwegian CDS has 30 percent monosyllables, 49 percent disyllables, and 22 percent polysyllables, whereas Norwegian ADS contains 25 percent monosyllables, 41 percent disyllables, and 34 percent polysyllables. The proportion of monosyllables is higher in target words and actual productions in children's speech than in content words in CDS, and higher in CDS than in ADS, whereas polysyllables are more common in ADS than in CDS, and very scarce among the children's word types. A chi-square test confirms that these differences are significant ($\chi^2(4, N = 1,812) = 59.99, p < 0.001$).

DISCUSSION

On the basis of cross-linguistic data from CDI reports, children's speech samples, child directed speech, and adult directed speech we have examined phonological aspects of the child's emergent lexicon. In accordance with previous literature, we found that children speaking English, Italian, Swedish, Danish, and Norwegian have an affinity for word initial bilabials, but that their proportions of mono-, di-, and polysyllables vary. The Norwegian speaking children used more word initial bilabials than adults, as well as a higher proportion of monosyllables, and the parents accommodated when speaking to their children in these two respects. We also found a higher proportion of monosyllables among the target words in the Norwegian CDI data than among the target words in the spontaneous speech data. This difference may be a coincidence, but it could also stem from differences in the age range and vocabulary sizes of the children in the two data sets: The CDI analyses were based on children aged 0;8-1;8, most of whom had a vocabulary smaller than 50 words (Simonsen et al., 2014), whereas the four children from the Garmann corpus were analyzed at the 50 word stage. Thus, the children in these two data sets may be at different stages in their lexical and phonological development.

Our first hypothesis was that the high proportion of initial bilabials in early target words and productions is a property of early words independent of language, and that we would find this pattern also for Norwegian children. This hypothesis was confirmed by analyses of Norwegian, Danish, Swedish, English, and Italian CDI results as well as analyses of targets and actual productions in spontaneous speech in Norwegian. Gayraud and Kern (2007) and Wehberg et al. (2007) claim that the proportion of initial bilabials is high in the list of first target words in Danish and French, and Wehberg et al. (2007) compare their results to similar results for American-English. Since we found no significant cross-linguistic differences in the proportions of initial bilabials in children's target words across Danish, Swedish, English, Italian, and Norwegian, a high proportion of initial bilabials may be a cross-linguistic characteristic of children's speech, at least between these five languages. Actually, this may be an even more universal tendency, as de Boysson-Bardies and Vihman (1991) compared French, English and Swedish with Japanese, and found high proportions of initial bilabials in children's target words in all these languages including the non-Indo-European language Japanese.[4]

The similarities in proportions of word initial bilabials cannot simply reflect that the CDI forms are so similar from language to language; results from spontaneous speech both in this study and in de Boysson-Bardies and Vihman (1991) support the suggestion that a high proportion of initial bilabials is a cross-linguistic characteristic of children's speech as opposed to being input-related. de Boysson-Bardies and Vihman (1991, p. 312–313) argue that the lip closure associated with these sounds is a visual cue, which gives the child an opportunity to observe how they are produced. McCune and Vihman (2001) add that children have better control over the lips and the jaw than over the various parts of the tongue and that the proprioceptive feedback from the lips is easier to interpret than the feedback of lingual consonants, two factors that make bilabials easier to produce than other consonants. MacNeilage and Davis (2000) and Davis et al. (2002) have promoted the argument of motoric control, connecting children's first words to the evolution of language and claiming that the syllabic structure of language develops from the repetitive movements of the jaw and the tongue during the process of digestion. The preference for initial labials may have to do with the relative ease of moving the jaw and the lips compared to moving the tongue. Mulford (1988), on the other hand, accentuates the importance of visual input for the labial cross-linguistic tendency, reporting that blind children produce a lower proportion of labials than sighted children. As our data cannot inform us in this debate, we simply notice that several factors may work in favor of the same result.

Our second hypothesis was that word length in syllables in early target words and productions varies across languages, and that Norwegian children's first words would be balanced between mono- and disyllabic words, similarly to previous findings from Swedish children (Vihman and Croft, 2007). This hypothesis was confirmed by our cross-linguistic comparison of CDI data, as well as by results from analyses of spontaneous speech. We found cross-linguistic differences in the lists of first target words; English and Swedish are balanced, like Norwegian, but Danish is skewed toward monosyllables and Italian is skewed toward di- and polysyllables. Thus, according to both CDI data and analyses of targets and actual productions in spontaneous speech, Norwegian children follow the Swedish pattern, producing a balanced proportion of mono- and disyllabic target words. It seems like Danish, then, is the odd language out among the Scandinavian languages.

From the literature, we expected English and Danish to have particularly high proportions of monosyllables, and Italian to have a particularly high proportion of polysyllables. Our results differ somewhat from our expectations: Even though English children seem to produce more monosyllabic target words than Swedish children, this difference was not significant. It is therefore less clear whether English is so different from other languages when it comes to word length in syllables, at least when we look at CDI data. It would be interesting to compare our results with CDI data from German and Dutch, to see whether Germanic languages are different from other languages in not having dominantly disyllables among the early target words (see Vihman and Croft, 2007, p. 687).

Our third hypothesis was that Norwegian adults would produce fewer words with initial bilabials and longer words than Norwegian infants, but make adaptations with respect to both properties in CDS. The corpus analyses support this hypothesis, as the proportions in CDS were between children's speech and

[4]Note that the figures in de Boysson-Bardies and Vihman (1991) are not directly comparable with ours, as they excluded words with an initial vowel, glide or /h/ in their calculations.

ADS with respect to both properties investigated here. It is well known that adults accommodate their speech in many ways when addressing young children (Snow, 1972; Cruttenden, 1994; Englund, 2005; Englund and Behne, 2005). Our results add to this literature, suggesting that adults accommodate in phonological detail to children when they talk to them (see also Cruttenden, 1994), using fewer polysyllabic words and more words with an initial bilabial than when talking to other adults. This could be a consequence of the children's influence on the conversation topics; *ball* 'ball,' *bil* 'car,' and *banan* 'banana' are all early words that children might want to talk about.

The analyses of Norwegian spontaneous speech indicate that children target fewer polysyllables than adults do. This is similar to the results found in Gayraud and Kern (2007), which show that French 2-year-olds target mono- and disyllables, but not polysyllables, whereas polysyllables first appear at 30 months, and become a substantial category at 46 months. They suggest that this may be due to a lower working memory capacity in younger children. The literature does indeed link the working memory capacity to language acquisition in both children and adults (Baddeley, 2003). A study of 4- and 5-year olds showed that it was easier to remember mono- and disyllabic non-words than polysyllabic non-words, but that disyllables were easier to remember than monosyllables (Gathercole and Baddeley, 1989).

Our analyses of spontaneous speech revealed that the group results mask individual differences for proportions of word initial bilabials and proportions of monosyllables, although to different degrees. Regarding initial bilabials in production, the children ranged from 8% (Johanna) to 51% (Iben). The proportions of monosyllables in their productions ranged from 33% (Johanna) to 65% (Olav). This suggests that children's actual productions might be more affected by their individual phonological preferences, as expressed in, e.g., templates (Vihman and Croft, 2007), than by the adult targets. To learn more about children's phonology, then, we need to take a closer look at individual differences in their actual productions.

CONCLUSION

We set out to investigate one possibly language-dependent property of children's early words, word length in syllables, and one possibly language-independent property, the proportion of word-initial bilabials, investigating child language data from five different languages and comparing children's and adults' speech for Norwegian. We think that our results and the following discussion support the view that children—at least across the languages investigated here—have an affinity toward words with initial bilabials, possibly related to visual cues, proprioception, and motoric control. The length of the words they acquire, on the other hand, depends more on properties of the ambient language. However, Norwegian children appear to overuse not only words with initial bilabials, but also monosyllables, when compared to CDS and ADS. Thus, our findings suggest that although the available target words in the ambient language influence the word length of early target words, factors such as memory capacity may still bias young children toward shorter words.

DATA AVAILABILITY

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The study was carried out in accordance with the recommendations of the Norwegian Center for Research Data. The parents who filled in CDI questionnaires for their children gave written informed consents on behalf of themselves and the children involved, and so did the parents and any other adult involved in child-parent interaction. Each of these two projects were approved by the Norwegian Center for Research Data.

AUTHOR CONTRIBUTIONS

NGG initiated the project and contributed with child language speech data. HGS and KEK contributed with CDI data. PH and NGG analyzed the speech data together, and PH made decisions on how to study the Norwegian CDI data, carried out the statistical analyses, and created the figures. All authors were involved in writing all parts of the manuscript, often simultaneously. All authors have approved of the final version of the manuscript.

FUNDING

This work was partly supported by the Research Council of Norway through its Centers of Excellence funding scheme, project number 223265, as well as an individual research project, project number 197337. We also wish to thank the Department of Linguistics and Scandinavian Studies, University of Oslo, and the Department of Early Childhood Education, OsloMet – Oslo Metropolitan University for financial support of the transcription of data.

ACKNOWLEDGMENTS

In addition to the two first authors, Anna Sara Romøren, Eirik Olsen, Elisabeth Holm, Ingeborg Sophie Ribu and Hannah Bergh-Johnsen contributed to transcription of the Garmann corpus. We wish to thank the audiences at the conference ELA 2012 at Université de Lyon, and several NorPhLex workshops for their valuable comments, among whom Marilyn Vihman, Tamar Keren-Portnoy and Dorthe Bleses have been particularly willing to discuss topics related to this paper. Yvan Rose and Greg Hedlund deserve thanks for all their help regarding the use of Phon. Finally, we thank LouAnn Gerken and Marilyn Vihman for insightful comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2019.00010/full#supplementary-material

REFERENCES

Baddeley, A. (2003). Working memory and language: an overview. *J. Commun. Disord.* 36, 189–208. doi: 10.1016/S0021-9924(03)00019-4

Basbøll, H. (2005). *The Phonology of Danish.* Oxford: Oxford University Press.

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T. O., et al. (2008). Early vocabulary development in Danish and other languages: a CDI-based comparison. *J. Child Lang.* 35, 619–650. doi: 10.1017/S0305000908008714

Braginsky, M., Yurovsky, D., Marchman, V. A., and Frank, M. C. (2016). "From uh-oh to tomorrow: predicting age of acquisition for early words across languages," in *Proceedings of the 38th annual conference of the Cognitive Science Society* (Philadelphia, PA), 1691–1696.

Bybee, J. (2010). *Language, Usage and Cognition*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511750526

Caselli, M. C., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., et al. (1995). A cross-linguistic study of early lexical development. *Cogn. Dev.* 10, 159–199. doi: 10.1016/0885-2014(95)90008-X

Cruttenden, A. (1994). "Phonetic and prosodic aspects of Baby Talk," in *Input and Interaction in Language Acquisition,* eds C. Gallaway, and B. J. Richards (New York, NY: Cambridge University Press), 135–152.

Davis, B. L., MacNeilage, P. F., and Matyear, C. L. (2002). Acquisition of serial complexity in speech production: a comparison of phonetic and phonological approaches to first word production. *Phonetica* 59, 75–107. doi: 10.1159/000066065

de Boysson-Bardies, B., and Vihman, M. M. (1991). Adaptation to language: evidence from babbling and first words in four languages. *Language* 1991, 297–319. doi: 10.1353/lan.1991.0045

Englund, K. T. (2005). Voice onset time in infant directed speech over the first six months. *First Lang.* 25, 219–234. doi: 10.1177/0142723705050286

Englund, K. T., and Behne, D. M. (2005). Infant directed speech in natural interaction—Norwegian vowel quantity and quality. *J Psycholinguistic Res.* 34, 259–280. doi: 10.1007/s10936-005-3640-7

Eriksson, M., and Berglund, E. (1999). Swedish early communicative development inventories: words and gestures 1. *First Lang.* 19, 55–90. doi: 10.1177/014272379901905503

Fenson, L., Dale, P. S., Hartung, J., Reilly, J., Bates, E., Pethick, S. J., et al. (2007). *MacArthur-Bates Communicative Development Inventories: User's Guide and Technical Manual.* San Diego, CA: Thomson.

Frank, M. C., Braginsky, M., Yurovsky, D., and Marchman, V. A. (2017). Wordbank: an open repository for developmental vocabulary data. *J. Child Lang.* 44, 677–694. doi: 10.1017/S0305000916000209

Garmann, N. G., and Torkildsen, J. V. K. (2017). "Barns språkutvikling de tre første årene," in *Helt fabelaktig! Festskrift til Hanne Gram Simonsen på 70-årsdagen,* eds H.-O. Enger, M. I. N. Knoph, K. E. Kristoffersen, and M. Lind (Oslo: Novus), 45–64.

Garmann, N. G. (2016). *The Garmann Corpus*. CHILDES/Phon corpus. doi: 10.21415/T5P59D

Gathercole, S. E., and Baddeley, A. D. (1989). Evaluation of the role of phonological STM the development of vocabulary in children: a longitudinal study. *J Memory Lang.* 28, 200–213. doi: 10.1016/0749-596X(89)90044-2

Gayraud, F., and Kern, S. (2007): Caractéristiques phonologiques des noms en function de l'âge d'acquisition. *Enfance* 59, 324–338.doi: 10.3917/enf.594.0324

Hagen, K., and Simonsen, H. G. (2014). "Two Norwegian speech corpora: NoTa–Oslo and TAUS," in *The Oxford Handbook of Corpus Phonology,* eds J. Durand, U. Gut, and G. Kristoffersen (Oxford: Oxford University Press), 498–508.

Haman, E., Łuniewska, M., Hansen, P., Simonsen, H. G., Chiat, S., Bjekic, J., et al. (2017). Noun and verb knowledge in monolingual preschool children across 17 languages: data from Cross-linguistic Lexical Tasks (CLT). *Clin. Ling. Phonet.* 31, 818–843. doi: 10.1080/02699206.2017.1308553

Haman, E., Łuniewska, M., and Pomiechowska, B. (2015). "Designing Cross-linguistic Lexical Tasks (CLTs) for bilingual preschool children," in *Methods for assessing multilingual children: Disentangling bilingualism from Language Impairment*, eds S. Armon-Lotem, J. d. Jong, and N. Meir (Bristol, UK: Multilingual Matters), 196–240.

Hansen, P., Simonsen, H. G., Łuniewska, M., and Haman, E. (2017). Validating the psycholinguistic aspects of LITMUS-CLT: evidence from Polish and Norwegian. *Clin. Ling. Phonet.* 31, 910–930. doi: 10.1080/02699206.2017.1307455

Keren-Portnoy, T., Majorano, M., and Vihman, M. M. (2009): From phonetics to phonology: the emergence of first words in Italian. *J. Child Lang.* 36, 1–33. doi: 10.1017/S0305000908008933

Kristoffersen, K. E., and Simonsen, H. G. (2012). *Tidlig språkutvikling hos norske barn. MacArthur-Bates foreldrerapport for kommunikativ utvikling*. Oslo: Novus Forlag.

Kristoffersen, K. E., Simonsen, H. G., Bleses, D., Wehberg, S., Jørgensen, R. N., Eiesland, E.-A., et al. (2013). The use of the Internet in collecting CDI data – an example from Norway. *J. Child Lang.* 40, 567–585. doi: 10.1017/S0305000912000153

Law, J., and Roy, P. (2008). Parental report of infant language skills: a review of the development and application of the Communicative Development Inventories. *Child Adolesc. Mental Health* 13, 198–206. doi: 10.1111/j.1475-3588.2008.00503.x

MacNeilage, P. F., and Davis, B. L. (2000). On the origin of internal structure of word forms. *Science* 288, 527–531. doi: 10.1126/science.288.5465.527

Majorano, M., Vihman, M. M., and DePaolis, R. A. (2014). The relationship between infants' production experience and their processing of speech. *Lang. Learn. Dev.* 10, 179–204. doi: 10.1080/15475441.2013.829740

McCune, L., and Vihman, M. M. (2001). Early phonetic and lexical development: a productivity approach. *J. Speech Lang. Hear. Res.* 44:670. doi: 10.1044/1092-4388(2001/054)

Mulford, R. (1988). "First words of the blind child," in *The Emergent Lexicon: The Child's Development of a Linguistic Vocabulary,* eds M. D. Smith and J. L. Locke (San Diego, CA: Academic Press), 293–338.

Peña, E. D. (2007). Lost in translation: methodological considerations in cross-cultural research. *Child Dev.* 78, 1255–1264. doi: 10.1111/j.1467-8624.2007.01064.x

Rose, Y., MacWhinney, B., Byrne, R., Hedlund, G., Maddocks, K., O'Brien, P., et al. (2006). "Introducing Phon: a software solution for the study of phonological acquisition," in *Proceedings of the 30th Annual Boston University Conference on Language Development*, eds D. Bamman, T. Magnitskaia, and C. Zaller (Somerville, MA: Cascadilla Press), 489–500.

Simonsen, H. G., Kristoffersen, K. E., Bleses, D., Wehberg, S., and Jørgensen, R. N. (2014). The Norwegian Communicative Development Inventories – reliability, main developmental trends and gender differences. *First Lang.* 34, 3–23. doi: 10.1177/0142723713510997

Snow, C. E. (1972). Mothers' speech to children learning language. *Child Dev*. 1972, 549–565. doi: 10.2307/1127555

University of Oslo (2013). *Norsk Talespråkskorpus – Oslodelen (NoTa–Oslo) [The Norwegian Speech Corpus – The Oslo Part*]. The Text Laboratory, Department of Linguistics and Scandinavian studies, University of Oslo. Available online at: http://tekstlab.uio.no/nota/oslo

Vihman, M., and Croft, W. (2007). Phonological development: toward a 'radical' templatic phonology. *Linguistics* 45, 683–725. doi: 10.1515/LING.2007.021

Vihman, M. M., Kay, E., de Boysson-Bardies, B., Durand, C., and Sundberg, U. (1994). External sources of individual differences? A cross-linguistic analysis of the phonetics of mothers' speech to 1-yr-old children. *Dev. Psychol.* 30:651. doi: 10.1037/0012-1649.30.5.651

Wehberg, S., Vach, W., Bleses, D., Thomsen, P., Madsen, T. O., and Basbøll, H. (2007). Danish children's first words: analysing longitudinal data based on monthly CDI parental reports. *First Lang.* 27, 361–383. doi: 10.1177/0142723707081723

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX: EXCERPTION OF CONTENT WORDS IN ADULT SPEECH

Table A1 gives a summary of content vs. function words, following Vihman et al. (1994). Different stylistic forms of the same word were disregarded, e.g. *dansa* and *danset* for past tense of *danse* 'dance'. (This was not considered in Vihman et al., 1994). Inflected forms of the same lexeme are regarded as separate word types. Verb particles were treated as part of the verb, which means that verb + particle sequences are also considered separate word types, e.g. *ha på* 'have' + 'on' = 'wear', *ha på deg* 'have' + 'on' + 'you' = 'wear' and *ha det på* 'have' 'it' 'on' = 'wear it'. The definition of content words in Vihman et al. (1994) is wide – 54 percent of the word tokens and 84 percent of the word types in our sample were regarded as content words.

It turned out to be difficult to decide which sequences of verb + preposition/adverb were to be interpreted as verb + particle, and which were to be interpreted as verb followed by a prepositional phrase or an adverb. The following criteria were used:

1. If the verb + preposition/adverb does not take a complement, the sequence is verb + particle. If there is an implicit complement in the utterance, or the sequence does take a complement, supplementary criteria were used:
2. If the verb normally has tonal accent 1, but changes to tonal accent 2 when followed by a preposition or an adverb, the sequence is interpreted as a verb + particle. This is a systematic distinction in Urban East Norwegian, for example: ^{1}*ta på kjolen* which means 'touch the dress' and ^{2}*ta på kjolen* which means 'put on the dress'. This example also illustrates how the meaning is different when the preposition or adverb has grammaticized to a particle and where the word has kept its grammatical meaning.
3. If the preposition or adverb is accented, it cannot be a particle. From this follows that normally the particle is monosyllabic: for example as *med* in *vi kan ta med krakken* 'we can bring the stool'. However, particles may also be polysyllabic as in *se etter* 'look after' and *komme tilbake* 'return': To count as a particle, the preposition or adverb must then be unaccented and the verb + particle unit must have tonal accent 2.
4. Occasionally, there may be an object between the verb and the particle, e.g. *putte den inn* 'put it in'. The object will then be both short and unaccented. All variants of PUTTE + X + INN were counted as exemplars of one type. If the word(s) between the verb and the preposition/adverb is long and/or accented, the preposition/adverb is not considered to be a particle in a unit, e.g. *ser litt annerledes ut* 'looks a bit different', and *sitter helt bom fast* 'is totally stuck', although both *ser ut* 'looks' *and sitter fast* 'is stuck' are verb + particle units. Another type of verbs that may be counted as one or two is verbs followed by a reflexive pronoun. These pronouns are treated as part of the verb, e.g. *lene seg* 'to lean'.

TABLE A1 | Criteria for content and function words, based on Vihman et al. (1994).

Content words	Function words
Nouns	
Main verbs	Copulas. Auxiliaries (e.g., *vil* 'will' and *skal* 'shall'), and catenatives, grammaticized verbs as first verbs in a verb phrase, e.g., *la* in *la den være* 'let I be.'
Adjectives and adverbs	Pragmatic particles (*Hva skal du,* ***'a*** *(da)?'* Where are you going, **then**?')
Conventional interjections (e.g., *ops* 'oups')	
Onomatopoeia (e.g., *vov-vov* for 'dog')	Unconventional onomatopoeia and interjections (e.g., *bam* for 'bang,'), sound effects.
Simple formulaic routines (e.g., *værsågod* for 'here you are')	
	Articles, quantifiers
	Conjunctions, prepositions
	Pro-forms, question words
Locatives as true deictics (e.g., *Se der borte* 'Look over there')	Locatives as introducers or dummy forms (e.g., *Se her, nå skal vi ta på jakka.* 'Here, let's put on a jacket')

frontiers in Psychology

SYSTEMATIC REVIEW
published: 04 April 2019
doi: 10.3389/fpsyg.2019.00445

What Influences Language Impairment in Bilingual Aphasia? A Meta-Analytic Review

Ekaterina Kuzmina[1], Mira Goral[1,2], Monica Norvik[1,3,4] and Brendan S. Weekes[5,6,7]*

[1] Faculty of Humanities, Center for Multilingualism in Society Across the Lifespan, University of Oslo, Oslo, Norway, [2] The Graduate Center and Lehman College, City University of New York, New York, NY, United States, [3] Department of Speech and Language Disorders, Statped, Oslo, Norway, [4] Department of Language and Literature, Norwegian University of Science and Technology, Trondheim, Norway, [5] Melbourne Graduate School of Education, University of Melbourne, Melbourne, VIC, Australia, [6] Laboratory for Communication Science, Faculty of Education, The University of Hong Kong, Pokfulam, Hong Kong, [7] State Key Laboratory for Brain and Cognitive Sciences, The University of Hong Kong, Pokfulam, Hong Kong

Patterns of language impairment in multilingual speakers with post-stroke aphasia are diverse: in some cases the language deficits are parallel, that is, all languages are impaired relatively equally, whereas in other cases deficits are differential, that is, one language is more impaired than the other(s). This diversity stems from the intricate structure of the multilingual language system, which is shaped by a complex interplay of influencing factors, such as age of language acquisition, frequency of language use, premorbid proficiency, and linguistic similarity between one's languages. Previous theoretical reviews and empirical studies shed some light on these factors, however no clear answers have been provided. The goals of this review were to provide a timely update on the increasing number of reported cases in the last decade and to offer a systematic analysis of the potentially influencing variables. One hundred and thirty cases from 65 studies were included in the present systematic review and effect sizes from 119 cases were used in the meta-analysis. Our analysis revealed better performance in L1 compared to L2 in the whole sample of bilingual speakers with post-stroke aphasia. However, the magnitude of this difference was influenced by whether L2 was learned early in childhood or later: those who learned L2 before 7 years of age showed comparable performance in both of their languages contrary to the bilinguals who learned L2 after 7 years of age and showed better performance in L1 compared to L2. These robust findings were moderated mildly by premorbid proficiency and frequency of use. Finally, linguistic similarity did not appear to influence the magnitude of the difference in performance between L1 and L2. Our findings from the early bilingual subgroup were in line with the previous reviews which included mostly balanced early bilinguals performing comparably in both languages. Our findings from the late bilingual subgroup stressed the primacy of L1 and the importance of age of L2 learning. In addition, the evidence from the present review provides support for theories emphasizing the role of premorbid proficiency and language use in language impairment patterns in bilingual aphasia.

Keywords: bilingual aphasia, stroke, linguistic similarity, AoA, premorbid proficiency, language use, meta-analysis

OPEN ACCESS

Edited by:
Gonia Jarema,
Université de Montréal, Canada

Reviewed by:
Arturo Hernandez,
University of Houston, United States
Matti Laine,
Åbo Akademi University, Finland

****Correspondence:***
Ekaterina Kuzmina
rina.kuzmina@gmail.com

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *05 November 2018*
Accepted: *13 February 2019*
Published: *04 April 2019*

Citation:
Kuzmina E, Goral M, Norvik M and Weekes BS (2019) What Influences Language Impairment in Bilingual Aphasia? A Meta-Analytic Review. Front. Psychol. 10:445. doi: 10.3389/fpsyg.2019.00445

INTRODUCTION

Aphasia describes a multitude of acquired language impairment resulting from brain injury, most often but not exclusively following a stroke. Bilinguals are individuals who use more than one language on a regular basis (Grosjean, 2013). Reports of individuals with bilingual aphasia have emerged as an important constraint on theories of the neurobiology of language (Gollan and Kroll, 2001; Ullman, 2001; Abutalebi et al., 2009; Miozzo et al., 2010; Weekes, 2010). Studies of bilingual aphasia began with anecdotal case studies reported by Ribot (1882) and Pitres (1895/1983). However, the wider theoretical implications of these cases are only more obvious today with the advent of sophisticated models of bilingual language processing. Two enduring questions in the field are whether a first-acquired language (L1) is less vulnerable to brain damage compared with later-learned languages (L2), and whether a language that is used more often premorbidly can be privileged after injury. Ribot's law holds that earlier acquired memories (including linguistic) are more resistant to brain damage whereas Pitres' law assumes that the premorbidly dominant language will be less vulnerable, independent of the age of acquisition (AoA) of that language. A related question is whether the cognitive and neural representations for L1 and L2 are shared or depend on different cognitive and neural mechanisms (e.g., Chee et al., 1999, 2000; Abutalebi et al., 2001; Ullman, 2001; Green, 2003; Perani and Abutalebi, 2005; Giussani et al., 2007). In our view, answers to these questions can be revealing for theories of the neurobiology of language (e.g., Libben, 2017) as well as for the design of intervention for language impairments in multilingual speakers in a variety of contexts, including immigrants and refugees across the globe (Pot et al., 2018).

The evidence emerging from studies of bilingual individuals who are recovering language function after a stroke shows that both early acquisition and premorbid language dominance contribute to language recovery and should constrain therapy (Lorenzen and Murray, 2008; Faroqi-Shah et al., 2010; Knoph, 2013; Conner et al., 2018). In many instances, equivalent patterns of aphasia in all languages spoken premorbidly are assumed, an assumption that implies shared cognitive and neural representations for these languages. The shared bilingual neural substrate (SBNS) hypothesis specifically assumes that bilingual speakers who acquire L2 early in life have a common neural network with shared lexical-semantic and syntactic representations from each language in the brain (Miozzo et al., 2010; Costa et al., 2012; Nadeau, 2019). This assumption is compatible with cognitive neuropsychological models of typical bilingual language processing (Gollan and Kroll, 2001; Faroqi-Shah and Waked, 2010) and with the view that linguistic differences between languages spoken premorbidly do not matter. One prediction from these accounts is that brain damage from stroke will result in equivalent impairment for bilingual speakers in any two languages spoken premorbidly (Paradis, 2004; Weekes, 2010).

Methodological limitations in the sampling of multilingual people with aphasia reported in previous reviews, as detailed below, and the generally accepted view that L2 processing is moderated by AoA (for a review see Abutalebi, 2008), lead us to conjecture that language status (L1 vs. L2) would be a significant predictor of language impairment after stroke for bilingual speakers. According to the convergence hypothesis proposed by Green (2003), which is consistent with the SBNS, dissociations observed in bilingual speakers between L1 and L2 could reflect greater recruitment of cognitive resources assumed to be necessary to process an explicitly learned language (L2), rather than differential neuronal representations (see Chee et al., 1999, 2000; Ullman, 2001). Furthermore, the dominance of language use in the linguistic environment of a person with aphasia will have an impact upon the patterns of aphasia after stroke, according to Pitres' law (Goral et al., 2012; see also Gollan et al., 2015). Therefore, there is merit to explore the roles that AoA, premorbid language proficiency and use, as well as language similarity have on performance in bilingual speakers after a stroke.

The goal of the present meta-analysis is thus to examine what constrains language impairment following stroke in multilingual speakers, and specifically, to investigate whether AoA, premorbid language proficiency, use and exposure, as well as linguistic similarity between spoken languages determine reported patterns of aphasia in L1 and L2.

BACKGROUND

Decades of research show that language difficulties associated with aphasia are highly selective and can affect only one language modality (e.g., comprehension vs. production) or linguistic aspect (e.g., syntactic processing). Many persons living with aphasia are multilingual (Roberts and Kiran, 2007; Ansaldo and Saidi, 2014). When a multilingual speaker has aphasia following a stroke, the languages spoken premorbidly may show comparable or differential patterns of impairment (Paradis, 2004; Weekes, 2010). Differential patterns may manifest as greater impairment in one language compared to another, or as differences in the characteristics of aphasia. The reasons for differential impairments are less certain. Theories of differential language processing and of impaired mechanisms of language control have been put forward to account for the patterns observed (e.g., Ullman, 2001; Abutalebi and Green, 2008). Furthermore, research shows that AoA, premorbid language proficiency, use and exposure, as well as linguistic similarity between spoken languages influence patterns of differential impairment observed in multilingual aphasia (e.g., Fabbro, 2001; Paradis, 2001, 2004; Lorenzen and Murray, 2008; Goral et al., 2012, 2013).

PREVIOUS REVIEWS AND STUDIES

Previous reviews asked whether multilingual speakers with aphasia evidence comparable levels of language impairment in all languages spoken premorbidly. For example, Albert and Obler (1978) reviewed 108 cases of multilingual aphasia and found comparable distributions of parallel and non-parallel impairment among those who were early "compound" bilinguals and those who learned their L2 later in life. Their review demonstrated no dominant pattern of results supporting only Ribot's law or only Pitres' law, and that variables, such as age, age of language acquisition, and education influenced the outcome.

Paradis (2001) reviewed 132 cases published in the period from 1990 to 1999 and found that 81 cases (61%) showed parallel impairment in both languages ["when both (or all) languages are similarly impaired and restored at the same rate," p. 70], 24 (18%) had differential impairment ("impairment is of different degree in each language relative to premorbid mastery," p. 70), and the remainder was shared by 12 cases (9%) with blended impairment ("when patients systematically mix or blend features of their languages at any all levels of linguistic structure," p. 70), 9 cases (7%) with selective impairment ("when patients do not regain the use of one or more of their languages," p. 70), and 6 cases (5%) with successive impairment ("when one language does not begin to reappear until another has been maximally recovered," p. 70). It is important to notice that the distribution of the impairment pattern percentages in this review was highly influenced by the two relatively large group studies from which 99 cases (75%) were taken. In the first study by Junqué et al. (1995), impairment patterns of 50 early equally proficient Catalan-Spanish aphasic bilinguals with unequal premorbid frequency of language use were reported. In the second one by Vilariño et al. (1997), comparable impairment corresponding to premorbid proficiency was reported as the most frequent pattern based on the assessment of 49 early Galician-Spanish bilinguals with aphasia. Thus, the majority (75%) of the 132 cases included in the Paradis (2001) review were early, relatively balanced bilingual speakers of two closely related languages. Additionally, it is important to note that, firstly, the reviewed cases were of varying etiology (e.g., stroke, tumor), secondly, it was not systematically specified what language performance measures were used for assessment, and, finally, the criteria for making a decision about the comparability of impairments in both languages were not explicitly defined.

Fabbro (2001) used the Bilingual Aphasia Test (BAT, Paradis and Libben, 1987) to assess 20 Friulian-Italian early (L2 learned between 5 and 7 year) bilingual speakers with aphasia, who premorbidly used both languages on a regular basis and had a stroke from 1 to 96 months before the assessment. Premorbid proficiency of the participants was not directly assessed, the author allegedly assumed that all participants were equally proficient in both languages. According to the author's interpretation of the results, 13 participants (65%) had comparable impairments in both languages, 4 participants (20%) performed significantly worse in L2, and 3 participants (15%) performed significantly worse in L1 (We note that for one of these three last participants, p-value was 0.07 indicating the absence of significant differences.) The researcher concluded that these percentages were in line with the previous review by Paradis (2001). However, Fabbro's study included early balanced (having comparable premorbid proficiency in both languages) bilinguals only. Moreover, decisions about the difference between performance in L1 and L2 were made based on running significance tests separately for each of 20 participants, subjecting the results to a Type I error (overestimation of significant difference).

Other reviews have identified additional relevant variables. In their review, Lorenzen and Murray (2008) suggested that language similarity (proportion of cognates shared) was a significant constraint on language recovery in bilingual speakers after stroke. Ansaldo et al. (2008) argued, as others have earlier (see Paradis, 2004), that motivation impacts on recovery. Overall, extant reviews suggest that equivalent patterns of language impairment in bilingual aphasia are more common, but a large minority of cases do show differential or selective patterns of impairment. These reviews also highlight the variables that predict recovery in post-stroke bilingual aphasia: AoA, language proficiency, language use, and linguistic similarity.

AGE OF LANGUAGE ACQUISITION

AoA refers to the age at which people learn language. It has long been argued that words acquired at an early age are the ones that are most preserved in aphasia (Rochford and Williams, 1965; Brysbaert and Ellis, 2016; Bakhtiar et al., 2017) although experimental evidence has been mixed, with some later-learned words found to be more easily retrieved in some cases of aphasia (e.g., Goral et al., 2013). Much research has studied the question of whether languages that are learned later in childhood or in adulthood, as compared to early acquired first language(s), are organized or processed by different neural mechanisms (this discussion is beyond the scope of this paper but see Ullman, 2001; Birdsong, 2006; Abutalebi and Green, 2007 among others). In the literature on bilingual aphasia, most reports highlight the age in which the languages spoken were first acquired (e.g., for all 130 cases included in this review).

Whereas this question preoccupied early reviews (Albert and Obler, 1978; Junqué et al., 1995), relatively few recent studies of bilingual aphasia explicitly addressed the role of age of language learning on language impairment. Among those who did, Tschirren et al. (2011) found no evidence of differential performance in their late bilinguals, suggesting that late learning of L2 is not always an impediment after a stroke. They did, however, find that AoA had an impact on syntactic processing in the two languages. Other studies have found lower performance in a later-learned language than earlier-acquired ones despite pre-aphasia high levels of proficiency (e.g., Goral et al., 2006; Kiran and Iakupova, 2011; Kurland and Falcon, 2011). However, as noted by these authors, levels of premorbid proficiency in all languages spoken is difficult to assess. We address this issue next.

LANGUAGE PROFICIENCY

To determine language impairments in multilingual speakers with aphasia one needs to estimate their premorbid proficiency in these languages. However, premorbid language proficiency can only be estimated indirectly primarily via subjective ratings. Several questionnaires have been developed to elicit such ratings (Paradis and Libben, 1987; Muñoz and Marquardt, 2003; Kiran et al., 2010), but it has been demonstrated that self-ratings are not completely consistent with objective measures (Tomoschuk et al., 2018). Therefore, having no access to objective measures of proficiency prior to brain damage is a limitation. Recent studies have examined the notion that levels of language proficiency are highly related to levels of language exposure and use, suggesting that understanding patterns of language use could augment decisions about degree of language proficiency when only subjective measures are available (Kiran and Tuchtenhagen, 2005).

LANGUAGE USE AND EXPOSURE

Multilinguals are likely to attain and maintain high proficiency in languages they use regularly and frequently, especially if these languages are spoken in their living environment. When one language or more is not used, it could undergo processes of reduced activation and attrition (Köpke et al., 2007). Furthermore, living in the environment where one language is predominantly used can lead to inhibition of less used languages in immersed L2 learners (Linck et al., 2009). Thus, it can be assumed that the linguistic context at the time of the stroke can contribute to better perseverance and/or recovery of the relevant language in people with aphasia. In several studies, findings pointed to the role of the linguistic environment on the response to therapy (Goral et al., 2012, 2013), which are consistent with the importance of language context in addition to age of acquisition and language proficiency.

LINGUISTIC SIMILARITY

Differential performance between languages that are linguistically similar (e.g., Friulian and Italian) may be surprising but is in fact reported. Less expected is equivalent patterns of aphasia in languages that are linguistically different (e.g., Chinese and English). One reason for these reports is that manifestations of aphasia syndromes (e.g., agrammatism) are not possible in some languages and therefore similar patterns in linguistically different languages will not be observed (Weekes, 2010). Similarly, it is possible that different constraints that characterize linguistic systems (e.g., the depth of an orthography or complexity of morphology) will produce differential patterns of recovery (see Menn and Obler, 1990; Paradis, 2001; Weekes, 2005, 2012). When languages are similar in terms of their cognates (words that have a similar meaning and form in different languages), for example, Spanish and Catalan, linguistic distance is relatively small compared to languages without cognates, for example, Spanish and Mandarin. Despite this, selective impairments can be seen between linguistically similar languages.

Linguistic similarity has been associated with recovery in bilingual aphasia (Kohnert, 2004; Kendall et al., 2015). However, competition between cognates has also been observed (e.g., Kurland and Falcon, 2011). Linguistic similarity has been considered when testing differential impairment across languages. For example, Roberts and Deslauriers (1999) found that a group of 15 early balanced French-English bilinguals with aphasia were more accurate at naming pictures representing cognates compared to noncognates. As well, Goral et al. (2010) found cross-language effects from linguistic structures that were similar across languages but not for aspects that were different between languages of a trilingual speaker with aphasia. Similarly, Fabbro (2001) reported the most common error among Friulian-Italian bilinguals with aphasia when producing Friulian was pronoun omission, which is acceptable in many instances in Italian but ungrammatical in Friulian.

We note that the concept of linguistic similarity has been ill-defined in the literature. In the study of second or third language learning, one approach [Typological Primacy Model (TPM); Rothman, 2015] seeks to define language typology based on structural similarities and differences, rather than on the basis of language families and historical linguistics. In studies of bilingual language processing language similarities at the lexical level has been discussed with respect to the concept of cognates (e.g., Schepens et al., 2013). In most papers on bilingual aphasia no formal definition is offered (e.g., Ansaldo and Saidi, 2014).

CLINICAL RELEVANCE

Whether the first language of bilingual speakers who acquire aphasia is more likely to be better preserved, and the identification of the influencing variables that may moderate this outcome is not only interesting theoretically but is also critical to clinical practices. Language and communication assessment in aphasia may not reveal an accurate picture unless the individuals are assessed in all their languages and unless detailed information about their language history and use is obtained. Moreover, decisions about the language in which intervention is best conducted could be informed by evidence about relative degrees of impairments in the languages of the person with aphasia. There is a growing body of treatment studies that examine the effectiveness of intervention in aphasia depending on the language in which the treatment is delivered. Current findings are equivocal regarding the variables (premorbid proficiency, AoA, language use) that affect therapy outcome and cross-language generalization (e.g., Goral, 2012; Kiran et al., 2013a,b; Nadeau, 2019).

PRESENT STUDY

Aphasia is multidimensional and rarely presents as a pure syndrome in neurology (Caramazza, 1984; Caramazza and McCloskey, 1988; Nickels et al., 2011). Studies of bilingual speakers with aphasia—who are necessarily idiosyncratic in their language background—are ipso facto unique. It is therefore hardly surprising to find a majority of research in bilingual aphasia are case studies. Criticisms of case reports are longstanding, plentiful and still topical (Shallice, 1979; Caplan, 1988; McCloskey and Caramazza, 1988; Coltheart, 2017) and are not limited to the field of aphasia and have been usurped by the so called case-series approach (Schwartz and Dell, 2010; Lambon Ralph et al., 2011; Rapp, 2011). The defining quality of the case-series is the capacity to use patterns of covariance to understand underlying cognitive mechanisms, including key elements: a reasonable sample size suitable for identifying complex trends in idiosyncratic data; administration of a common set of cognitive tests; and open criteria for defining a sample motivated by theoretical questions or clinical and neuroanatomical criteria (Rapp, 2011). The single case is commonly associated with the universality assumption that characterizes "orthodox" cognitive neuropsychology (Caramazza, 1986), while the case-series approach is prima facie more compatible with the assumptions of "population thinking" (Bub, 2011). As Rapp (2011) notes despite the increase in "population thinking," little work has been done in aphasia to understand the extent and nature of individual variability with regard to the types of cognitive mechanisms commonly investigated in cognitive psychology and neuroscience.

In this meta-analysis we attempt to honor the variability presented in the case studies and the case-series, and at the same time extract patterns that transcend the variability and allow us to generalize from the existing literature. Given the theoretical questions raised about the neurobiology of language and the above-mentioned reports about potential predictors of impairments in aphasia in bilingual speakers, we aimed to answer the five following research questions:

1) Do bilingual speakers with post-stroke aphasia show a difference in performance between the language acquired first (L1) and the later learned language (L2)?
2) Are the possible differences between L1 and L2 of different magnitude between early bilinguals and late bilinguals? Does AoA as a continuous variable moderate the outcomes in the early and late bilingual subgroups separately?
3) Does premorbid language proficiency moderate the possible differences between L1 and L2?
4) Does frequency of language use moderate the possible differences between L1 and L2?
5) Does linguistic similarity between the languages spoken by the bilingual moderate the possible differences between L1 and L2?

METHODS

Literature Search

The following electronic databases were searched: PubMed, Science Direct, PsycINFO, CINAHL, TAYLOR, and FRANCIS Online. Five construct-related search terms (multilingual, bilingual, trilingual, quadrilingual, polyglot) and seven population-related search terms (aphasia, language disorder, language impairment, anomia, stroke, vascular, hemorrhage) were used. The search was limited to peer-reviewed papers published in the period from 2000 until 2018 and written in English. The search strings adapted for each database are reported in the **Supplementary Material**. First, titles of search hits were screened to define the relevance of a study to the review. Second, abstracts and method sections of results were screened for matching inclusion criteria.

Inclusion and Exclusion Criteria

Papers reporting behavioral accuracy data on language performance of multilingual persons with post-stroke aphasia were included for complete screening.

Two inclusion criteria related to participants were used. The first criterion was presence of aphasia resulting from a single cerebrovascular accident. Participants with aphasia of other etiologies (e.g., tumor, head injury, dementia) were excluded. The second criterion was the bilingual or multilingual status of participants. The categorization of participants as bilinguals or multilinguals by an author was used to decide whether to include participants into the review. Although variation in the definitions of bilingualism/multilingualism used by different authors can be assumed, all of the included participants could be described as persons who used more than one language to communicate on a regular basis in everyday life before the stroke. This was done to ensure that a participant had at least sufficient proficiency for everyday conversation prior to their stroke (B1 level according to the Common European Framework of Reference for Languages). Thus, the operational definition of bilingualism/multilingualism was primarily based on a criterion of premorbid language use (Grosjean, 1982). For participants whose performance was reported in several papers, information was taken from all of the papers, if the assessment time was equivalent. When the same person was described in multiple papers at different data points, the earliest performance was coded. Five studies reported data on more than two languages of the participants. For all of these cases, performance in L1 and the most frequently used language were extracted and analyzed. If several L2s were equally used, the earlier acquired language was chosen for the analysis.

Three inclusion criteria related to tests were also used. The first criterion was that a test should directly measure language performance (e.g., auditory syntactic comprehension, picture naming, reading aloud). Studies reporting performance only on tests indirectly measuring language performance (e.g., Color-Word Stroop) were not included. The second criterion was that reported performance was shown as correct responses out of the total number of tested items in various language tasks. Cases where accuracy was reported in percentages in a way that the total number of tested items in the task could not be estimated were excluded. Those studies where the total number of items used in the test was not reported, but a published test had this information (e.g., the Bilingual Aphasia Test) were included. The third criterion was the reported performance (accuracy and total number of items in the task) included data from more than one language. Cases were performance in only one language was reported were not included. After screening, 65 studies were included in the final dataset. **Figure 1** shows the details of the literature search and screening process with resulting number of studies.

Data Coding

Cases from the finally selected studies were coded according to the three study-related variables (first author or the first two authors, year of publication, first five words of the title), seven clinico-demographic case-related variables (gender, age in years at the time of assessment, years of education, month post onset at the assessment, type and severity of aphasia, lesion side), four language background variables (age of L2 acquisition, premorbid language proficiency, language use, linguistic similarity between languages), three test-related variables (test name, testing paradigm, language modality assessed by the test), and four

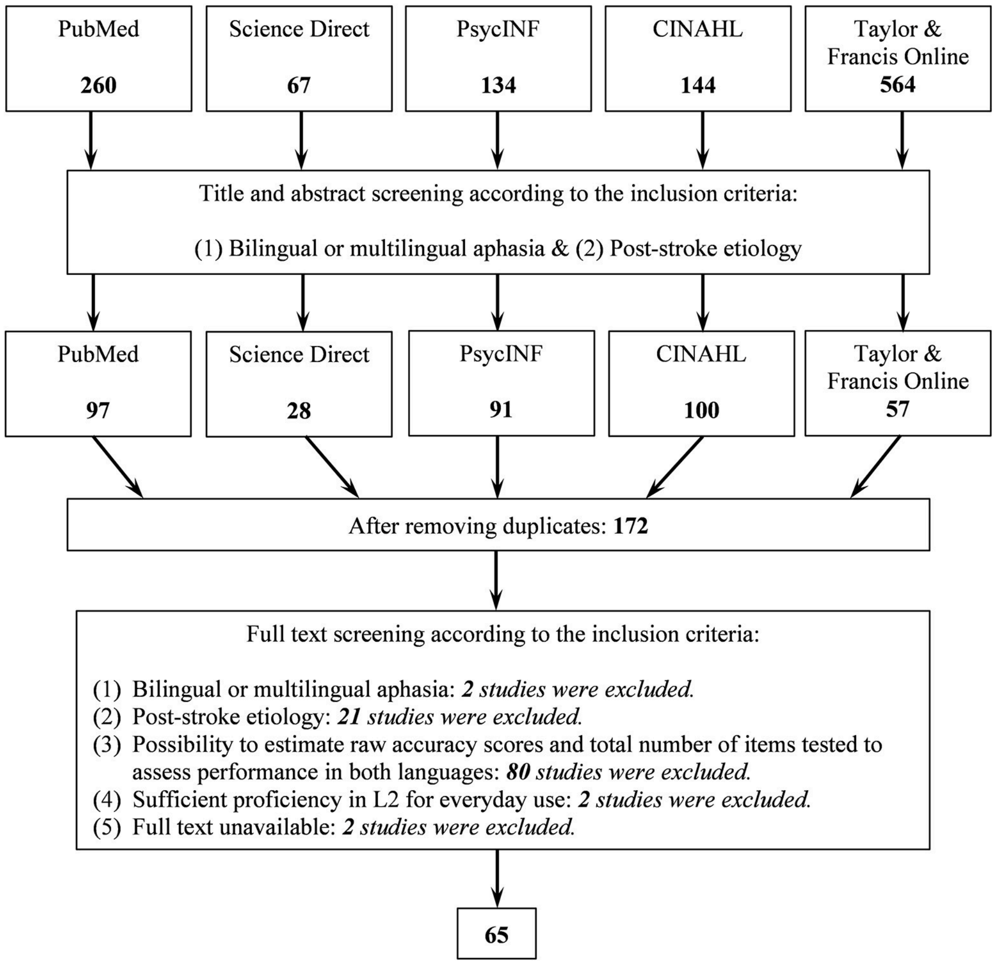


FIGURE 1 | Flowchart of the search process. Numbers show how many studies were included at each stage.

language performance variables (numbers of items correctly performed in L1 in a specific test, total number of items in the test used to assess L1, number of items correctly performed in L2 in a specific test, total number of items in the test used to assess L2).

Coding of several above mentioned variables requires elaboration. Age of L2 acquisition was coded either as a number if it was directly reported as such, or as a time period (i.e., early childhood, primary school, later than early childhood, high school, early adulthood, adulthood) if it was directly reported or could be inferred from case descriptions. Based on the age of L2 acquisition variable, we created an adjusted variable, where we transformed categorical labels into numbers according to the following criteria: early childhood = 3 year, primary school = 7 year, later than early childhood = 10 year, high school = 14 year, early adulthood = 20 year, adulthood = 25 year. This adjustment allowed us to perform the moderator analysis treating AoA as a continuous variable.

Language proficiency was coded using three levels, namely "higher in L1," "equal," "higher in L2" proficiency, based on the information from case descriptions. Language use was coded based on the information from the cases using also three levels, namely "more in L1," "equal," "more in L2" use. To assist the coding of the language use variable, we additionally coded the following variables: language used (1) to communicate with parents, (2) with other relatives, (3) with a partner, (4) with children, (5) with friends, (6) in school as an instruction language, (7) in further education as an instruction language, (8) as a subject of formal language classes, (9) at work, (10) for reading, (11) for writing, (12) to watch TV and listen to the radio, (13) based on a self-report, (14) for therapy, (15) in the environment as an official language. These variables were used to facilitate the decision on the language use variable.

The linguistic similarity variable was coded based on how far languages are located from each other in the language family classification in two ways (rather than using for instance the TPM Rothman, 2015, which is less feasible for a meta-analysis of this scope). Firstly, language pairs from different families (e.g., English is from Indo-European family and Chinese is from Sino-Tibetan family) were coded with the level "different," whereas all other pairs represented the "similar" level. Secondly,

TABLE 1 | Summary of the included modalities, testing paradigms, and tests.

AUDITORY COMPREHENSION MODALITY

1. Commands and Yes/No questions: *AAT:* Token test; *BAT:* Simple and semi-complex commands, Complex commands; *ILAT:* Commands; *MAST:* Yes/No questions; *WAB:* Commands, Yes/No questions.

2. Story or paragraph: *BAT:* Paragraph; *WAB:* Complex ideational material.

3. Auditory input to picture matching: *Authors' tasks:* Pointing - words, Pointing - sentences; *BAT*: Pointing - words, Auditory discrimination, Pointing - sentences; *BPVS:* Pointing - sentences; *CNL LSBA:* Lexical discrimination, Pointing - words; *ILAT:* Pointing - words; *PPVT:* Pointing - words; *PAL:* Pointing - words; *WAB:* Auditory discrimination, Pointing - words.

4. Syntactic grammaticality judgment: *Authors' task:* Grammaticality judgment; *BAT:* Grammaticality judgment; *CNL LSBA:* Grammaticality judgment.

5. Lexical decision: *Authors' task:* Lexical decision; *BAT:* Lexical decision; *CNL LSBA:* Lexical decision.

6. Semantic relationship judgment: *BAT:* Semantic acceptability, Semantic categories, Synonyms and antonyms, Semantic judgments.

7. Other measures: *Authors' task:* Auditory discrimination; *BAT:* Auditory comprehension, Auditory comprehension (pointing, semi-complex and complex commands), Auditory comprehension (pointing, semi-complex and complex commands, auditory discrimination), Sentence semantic violation judgment; *CAT:* Comprehension - words, sentences, and paragraph; *ILAT:* Phonemic analysis; *WAB:* Auditory comprehension, Auditory comprehension (Yes/No questions, word recognition, sequential commands).

ORAL PRODUCTION MODALITY

8. Confrontation picture naming: *AAT:* Naming; *Authors' task:* Naming - actions, Naming - objects; *BAT:* Naming - objects; *BNT:* Naming - objects; *CNL LSBA*: Naming - actions, Naming - objects; *ILAT:* Naming; *OANB:* Naming - objects and actions; *SWB:* Naming - objects; *WAB:* Naming; *Greek Action Test*: Naming - actions; *PALPA:* Naming.

9. Repetition: *AAT:* Repetition; *Authors' task:* Repetition - words and nonwords; *BAT:* Repetition - words and nonwords, Repetition - sentences; *CAT:* Repetition; *CNL LSBA:* Repetition - words, nonwords, and sentences; *PALPA:* Repetition; *WAB:* Repetition.

10. Responsive speech and sentence completion: *Authors' task:* Sentence completion; *CNL LSBA:* Sentence completion; *WAB:* Responsive speech, Sentence completion.

11. Sentence construction: *BAT:* Sentence construction; *CNL LSBA:* Sentence construction.

12. Semantic opposites: *BAT:* Semantic opposites.

13. Producing morphological derivatives: *BAT:* Morphological opposites; *CNL LSBA:* Morphological production, verb tense.

14. Spontaneous and semi-spontaneous production: *AAT:* Spontaneous production; *Authors' task:* Personal narrative in CIU (correct information units), Picture description in composite rubric scores; *BAT:* Picture description, Spontaneous speech; *BDAE:* Picture description; *CAT:* Picture description; *SPPA:* Picture description; *WAB:* Picture description, Narrative production.

OTHER MODALITIES

15. Reading aloud: *Authors' task:* Reading aloud - words and no words; *BAT:* Reading aloud - words, Reading aloud - sentences; *CAT:* Reading aloud; *WAB:* Reading aloud; *CNL LSBA:* Reading aloud - words, Reading aloud - nonwords.

16. Written comprehension: *Authors' task:* Visual lexical decision, Written word to picture matching; *BAT:* Reading comprehension - words, Reading comprehension - sentences, Reading comprehension - paragraph; *CAT:* Reading comprehension - words and sentences; *ILAT:* Reading comprehension - paragraph.

17. Written production: *AAT:* Writing; *BAT:* Copying, Writing to dictation - words, Writing to dictation - sentences; *CAT:* Copying; *CNL LSBA:* Writing to dictation; *PALPA:* Writing; *WAB:* Writing.

UNCATEGORIZED MEASURES

18. *AAT:* General comprehension; *BAT:* Semantics (semantic categories, synonyms and antonyms, semantic acceptability, semantic opposites); General comprehension, Total score; *MAST:* Total score.

AAT, Aachen Aphasia Test; BAT, Bilingual Aphasia Test; BDAE, Boston Diagnostic Aphasia Evaluation; BNT, Boston Naming Task; BPVS, British Picture Vocabulary Scale; CAT, Comprehensive Aphasia Test; CNL LSBA, Cognitive Neuropsychology Laboratory Language Screening Battery Action; ILAT, Israeli Loewenstein Aphasia Test; MAST, Mississippi Aphasia Screening Test; OANB, Object and Action Naming Battery; PAL, Psycholinguistic Assessment of Language; PALPA, Psycholinguistic Assessments of Language Processing in Aphasia; PPVT, Peabody Picture Vocabulary Test; SPPA, Sentence Production Program for Aphasia; SWB, Snodgrass and Vanderwart Battery; WAB, Western Aphasia Battery.

to make the coding of the linguistic similarity variable more precise, the three-level coding was applied: (1) language pairs from different families were coded "different" (e.g., English and Chinese), (2) language pairs which shared only the same family were coded "close" (e.g., German and Spanish), and finally, (3) language pairs which shared more than the same family were coded "very close" (i.e., English-Norwegian, Cantonese-Mandarin, Spanish-Catalan, Afrikaans-English, Malayalam-Kannada, English-Dutch, Yiddish-English, Balochi-Persian, Spanish-Italian, Italian-French, Kurdish-Persian, Spanish-French, Galician-Spanish).

Performance scores were recorded separately for each test (e.g., object naming, reading aloud words, syntactic auditory comprehension). **Table 1** represents the variety of the tests included in the analysis. For tests without a defined maximum score from the spontaneous and semi-spontaneous production testing paradigm, numbers of correct information units, and corresponding total numbers of units were used as measures.

Dealing With Heterogeneity in Measures

In the majority of the studies, participants were assessed with multiple tests. Firstly, scores from the individual tests were pooled together based on 18 testing paradigms summarized in **Table 1**. Then, scores from testing paradigms were pooled together based on the two main language modalities, namely auditory comprehension and oral production. Thus, seven testing paradigms (i.e., auditory comprehension of commands and yes/no questions, auditory comprehension of a story or paragraph, auditory based pointing, auditory syntactic grammaticality judgment, auditory lexical decision, auditory semantic relationship judgment, and other scores including sums of auditory comprehension related tests) were pooled together into auditory comprehension scores. Seven other testing paradigms (i.e., confrontation picture naming, repetition, responsive speech and sentence completion, sentence construction, oral production of semantic opposites, oral production of morphological derivatives, spontaneous and semi-spontaneous production) were pooled together into oral production scores. The other modalities category included three testing paradigms: reading aloud, written comprehension, and written production. Other tests which could not be categorized under these three modalities were kept separately. Finally, scores from auditory comprehension, oral production, other modalities, and uncategorized measures were pooled together to get the overall performance scores.

We performed correlational analysis to explore relationships between scores accumulated into the testing paradigms, scores pooled into the two main language modalities (auditory comprehension and oral production), and scores pooled into overall performance category (see **Table 2**). Spearman's correlation coefficients between the overall performance, total auditory comprehension, and total oral production scores varied from $rs = 0.57$ to $rs = 0.94$ suggesting moderate to very strong relationships. For the rest of the correlations, 79 (73%) varied from $rs = 0.61$ to $rs = 0.95$ indicating strong and very strong relationships, 21 (19%) varied from $rs = 0.40$ to $rs = 0.59$ indicating moderate relationships, and 8 (7%) correlation coefficients varied from $rs = 0.30$ to $rs = 0.39$ indicating weak relationships. Based on the results of this analysis, we concluded that the procedure of pooling scores from various test paradigms into the auditory comprehension and oral production modalities, as well as later pooling all available language performance scores into the overall performance category was justified.

Interrater Reliability

In the beginning of the coding stage, the authors coded three studies together and agreed on the coding criteria. Disagreements were resolved via discussion. Then the first author coded 40 studies, 62%, and the second and the third authors coded the remaining studies. Later we randomly selected 16 studies, 25%, which were coded by two authors. For the language use and premorbid proficiency variables, which often required decision making, all studies were coded by two authors and any discrepancies were resolved by discussion including three authors. The Cohen's kappa values suggested strong interrater agreement for both language use, $k = 0.807$, $p < 0.001$, and proficiency, $k = 0.818$, $p < 0.001$ variables (Fleiss et al., 2003).

Statistical Analysis

The *metafor* R package (Viechtbauer, 2010) was used for statistical analysis. To estimate effect sizes for the difference in performance between L1 and L2, we calculated risk ratios with the help of *escalc* function. According to the documentation of *escalc*, the argument RR provides logarithms of risk ratios, making them symmetric around zero as well as helping to decrease the positive skew in their distribution. The effect sizes in our sample are independent, because each effect size represents the difference in performance between L1 and L2 for a specific case.

First, we fitted random-effect models with the help of *rma* function to investigate whether there were differences in performance between L1 and L2 for the three types of scores: overall performance, auditory comprehension, and oral production. Then we performed the moderator analysis fitting mixed-effect models with the help of the same function to explore whether the possible difference in performance between L1 and L2 may be affected by the four variables of interest (i.e., early-late bilingual status, premorbid language proficiency, language use, and linguistic similarity). In addition to the moderator analysis on the early-late bilingual status variable, we analyzed whether AoA as a continuous variable moderates the outcomes in the early and late subgroups separately. The overall and moderator analyses were performed for the whole sample as well as for the early and late AoA subgroups, as well as separately for overall performance, auditory comprehension, and oral production scores.

Additionally, it was explored how participants' age at the time of assessment, years of education, and months post onset moderated the magnitude of the difference in performance between L1 and L2. The R scripts used for the analysis as well as the detailed report of the analysis are provided in **Supplementary Material**.

RESULTS

Data Screening

Three funnel plots, each showing distribution of effect sizes for overall performance, auditory comprehension, and oral production, were created to detect cases with immensely high standard errors (*SEs*) (see **Figure 2**). The standard error in the present analysis depended on the number of items used to assess a certain language modality: as the number of the tested items increases, SE gets smaller, and the precision gets higher. Based on visual examination of the funnel plots for overall performance, auditory comprehension, and oral production, the cut-off point was set at $SE = 0.3$. Thus, five, five, and 16 cases were removed for overall performance, auditory comprehension, and oral production scores, respectively. Given that large differences in performance between languages in the clinical population of persons with aphasia are meaningful and highly probable, we did not remove the data points with relatively large effect sizes. After deleting cases based on *SEs*, $log\ (RR) = -1.30$ had the largest absolute value among the datapoints from all three funnel plots. This value meant that in this case performance in L1 was 73% worse than in L2.

Descriptive Characteristics

A total of 65 peer-reviewed published studies, from which 130 cases were extracted, were included in the review. Given that the analysis we performed required having information on which

TABLE 2 | Correlations between the testing paradigms and the three types of scores used in the analysis.

Testing paradigms	L1						L2					
	Auditory comprehension total		Oral production total		Overall performance		Auditory comprehension total		Oral production total		Overall performance	
	rs	*n*	*rs*	*n*	*rs*	*n*	*rs*	*n*	*rs*	*n*	*rs*	*n*
AUDITORY COMPREHENSION MODALITY												
Commands and Yes/No questions	0.86**	48	0.59**	48	0.72**	48	0.88**	48	0.60**	48	0.75**	48
Story or paragraph	0.77**	23	0.61**	23	0.68**	23	0.82**	23	0.75**	23	0.83**	23
Auditory input to picture matching	0.83**	62	0.55**	54	0.74**	62	0.84**	62	0.62**	54	0.74**	62
Syntactic grammaticality judgment	0.85**	36	0.56**	27	0.74**	36	0.86**	36	0.72**	27	0.80**	36
Lexical decision	0.81**	38	0.67**	30	0.78**	38	0.85**	38	0.65**	30	0.79**	38
Semantic relationship judgment	0.84**	31	0.70**	31	0.78**	31	0.87**	31	0.68**	31	0.81**	31
Other	0.95**	21	0.71**	21	0.83**	21	0.90**	21	0.78**	21	0.84**	21
Auditory comprehension total			0.57**	83	0.80**	100			0.63**	83	0.80**	100
ORAL PRODUCTION MODALITY												
Confrontation picture naming	0.49**	79	0.89**	106	0.82**	106	0.59**	79	0.90**	106	0.84**	106
Repetition	0.47**	63	0.65**	64	0.61**	64	0.59**	63	0.72**	64	0.69**	64
Responsive speech and sentence completion	0.30	9	0.45	10	0.48	10	0.46	9	0.46	10	0.39	10
Sentence construction	0.79**	23	0.85**	24	0.90**	24	0.72**	23	0.86**	24	0.91**	24
Semantic opposites	0.82**	23	0.88**	25	0.89**	25	0.81**	23	0.85**	25	0.90**	25
Morphological derivates	0.87**	15	0.77**	15	0.80**	15	0.88**	15	0.85**	15	0.88**	15
Spontaneous and semi-spontaneous production	0.55*	17	0.73**	22	0.70**	22	0.48*	17	0.74**	22	0.69**	22
Oral production total					0.93**	113					0.94**	113
OTHER MODALITIES												
Reading aloud	0.40**	41	0.52**	33	0.65**	41	0.35*	41	0.30	33	0.55**	41
Written comprehension	0.83**	28	0.35	20	0.78**	28	0.61**	28	0.32	20	0.71**	28
Written production	0.38	23	0.53**	24	0.73**	24	0.50*	23	0.56**	24	0.69**	24
UNCATEGORIZED MEASURES												
	0.43*	27	0.75**	19	0.88**	27	0.30	27	0.52*	19	0.72**	27

* $p < 0.05$, ** $p < 0.01$.

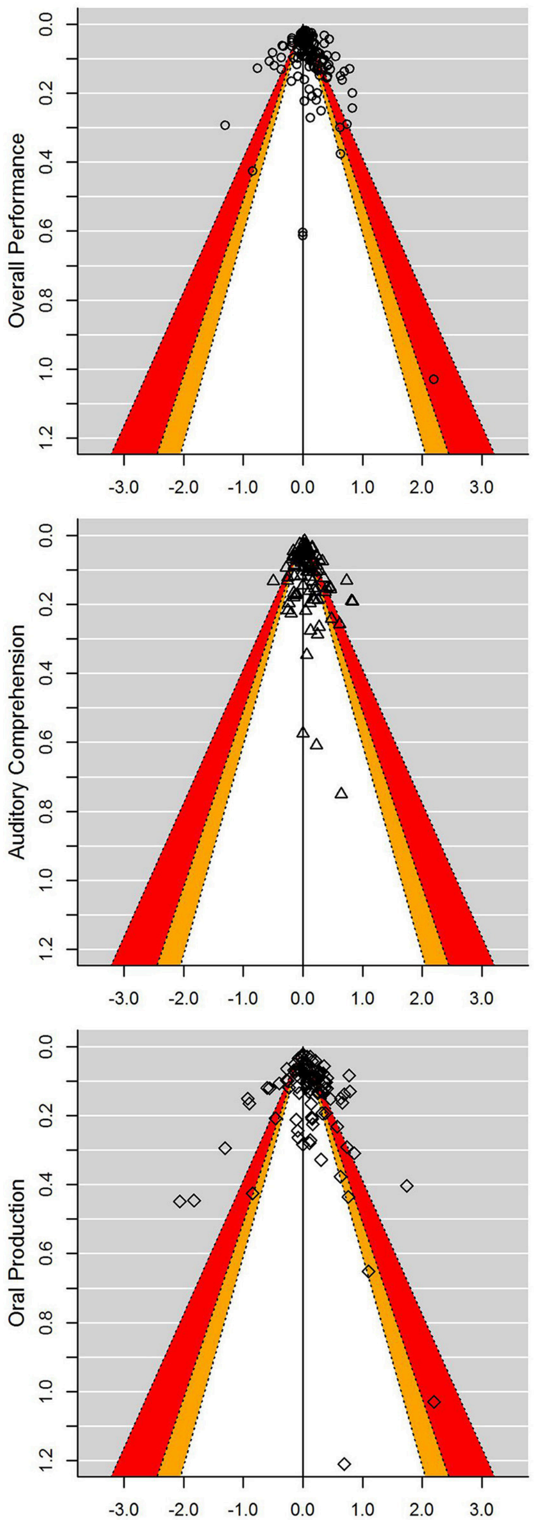


FIGURE 2 | Contour enhanced funnel plots for each of the three types of scores analyzed. Contours change shades at *p*-levels 0.1 (white), 0.05 (orange), and 0.01 (red). Logarithms of risk ratios are plotted against the *SEs*, and the reference line indicating the random-effects model estimates for each the three types of scores analyzed. Positive and negative abscissas represent better performance in L1 and L2, respectively.

language was acquired first, six cases representing simultaneous bilinguals who acquired both languages from the age of zero, were excluded from the sample.

Twenty seven (22%), 65 (52%), and 32 (26%) cases were taken from group ($n = 4$), multi-case ($n = 19$), and single-case studies ($n = 32$), respectively. Sixty two (50%) cases were extracted from studies with research questions unrelated to testing differences between the languages of multilingual people with aphasia ($n = 31$); the remaining 62 (50%) cases were extracted from studies with research questions related to testing differences between one's languages ($n = 24$). Detailed information about the cases is summarized in **Data Sheet 1** in **Supplementary Material**. Further analysis performed on the trimmed data showed that the study type (i.e., research question related vs. unrelated to testing L1/L2 differences) did not significantly moderate the outcomes for overall performance, $Q_M[1] = 2.89$, $p = 0.24$, auditory comprehension, $Q_M[1] = 0.21$, $p = 0.90$,or oral production, $Q_M[1] = 0.76$, $p = 0.68$.

Descriptive information on the demographic and clinical details of the sample used for the analysis as well as the early and late AoA subgroups is summarized in **Table 3**.

Language Status

After data trimming, the difference in performance between L1 and L2 was investigated using overall performance scores. We found a statistically significant effect size, $RR = 1.10$ [1.05, 1.15], $p < 0.0001$, Q_E [118] = 1025.14, suggesting that overall performance in L1 was on average 10% better than in L2 (see **Figure 3**). For auditory comprehension scores, we also found a statistically significant effect size, $RR = 1.06$ [1.02, 1.10], $p < 0.0001$, Q_E [90] = 363.41, suggesting that auditory comprehension in L1 was on average 6% better than in L2. Similarly, a statistically significant effect size, $RR = 1.10$ [1.03, 1.17], $p < 0.0001$, Q_E [90] = 686.25, was found for oral production scores suggesting that performance in L1was on average 10% better than in L2.

Age of Language Acquisition

Details of the moderator analysis (effects sizes, 95% CIs, and statistics of the moderator tests) are summarized in **Table 4**.

In the whole sample, AoA as a continuous variable moderated overall performance, $Q_M[1] = 8.84$, $p < 0.01$, and oral production, $Q_M[1] = 13.61$, $p < 0.001$, in the direction that as AoA increased, the magnitude of the L1 advantage (better performance in L1 compared to L2) increased. AoA as a continuous variable did not moderate auditory comprehension in the whole sample, $Q_M[1] = 1.65$, $p = 0.20$.

To decide on the cut-off point for making the early and late AoA subgroups, we visually explored the distribution of overall performance outcomes plotted against AoA as a continuous variable (see the plot in **Data Sheet 3**, p.53 in **Supplementary Material**). Based on this visual examination, 7 year appeared to be a reasonable cut-off point.

AoA status as a binary variable (early/late) significantly moderated overall performance, $Q_M[1] = 11.37$, $p < 0.001$, and oral production, $Q_M[1] = 8.85$, $p < 0.01$. Individuals

TABLE 3 | Demographic and clinical details of the whole sample and AoA subgroups.

Characteristics	Whole group			Early AoA subgroup (<7 year)			Late AoA subgroup (≥7 year)		
	$N = 119$			$n = 44$			$n = 75$		
	Mean	*SD*	Range	*Mean*	*SD*	Range	*Mean*	*SD*	Range
Age, year	58.5	14	17 - 91	52.9	14.2	17 - 84	61.8	12.9	33 - 91
Education, year	12.2	5.1	1 - 22	13.2	3.9	8 - 22	11.7	5.7	1 - 22
Female, % (*n*)	48% (57)			48% (21)			48% (36)		
Months post onset	28.3	14.9	1 - 53	28.3	15	2 - 53	28.3	15.0	1 - 52
AoA of L2, year	12.2	8.6	2.5 - 40	4.1	1.3	2.5 - 6	17.1	7.3	7 - 40
Lesion side: *n*	Left: 100; Right: 5; Both: 1; NA: 13			Left: 39; Right: 3; Both: 1; NA: 1			Left: 61; Right: 2; Both: 0; NA: 12		
Proficiency: *n*	L1: 27; Equal: 63; L2: 4; NA: 25			L1: 3; Equal: 30; L2: 4; NA: 7			L1: 24; Equal: 33; L2: 0; NA: 18		
Language use: *n*	L1: 27; Equal: 52; L2: 24; NA: 16			L1: 5; Equal: 21; L2: 13; NA: 5			L1: 22; Equal: 31; L2: 11; NA: 11		
Linguistic similarity, 2 levels: *n*	Similar: 90; Different: 29			Similar: 28; Different: 16			Similar: 62; Different: 13		
Linguistic similarity, 3 levels: *n*	Very close: 21; Close: 69; Different: 29			Very close: 7; Close: 21; Different: 16			Very close: 14; Close: 48; Different: 13		

who acquired L2 before 7 year showed a significantly smaller difference between L1 and L2 in overall performance, $RR = 1.00$ [0.93, 1.07], $p = 0.99$, compared to those who acquired L2 after 7 year, where performance in L1 was significantly better than in L2, $RR = 1.16$ [1.10, 1.23], $p < 0.0001$. Similarly for oral production, the early AoA subgroup showed a significantly smaller difference between L1 and L2, $RR = 0.97$ [0.88, 1.07], $p = 0.60$, compared to the late AoA subgroup, where performance in L1 was significantly better than in L2, $RR = 1.17$ [1.09, 1.26], $p < 0.0001$. The early-late bilingual status did not moderate auditory comprehension, $Q_M[1] = 2.59$, $p = 0.11$.

Additionally, we explored how AoA as a continuous variable moderated the outcomes in the early and late AoA subgroups separately. In the early AoA subgroup, AoA did not moderate overall performance, $Q_M[1] = 1.72$, $p = 0.19$, auditory comprehension, $Q_M[1] = 2.83$, $p = 0.09$, or oral production, $Q_M[1] = 0.06$, $p = 0.80$. In the late AoA subgroup, AoA moderated oral production, $Q_M[1] = 6.15$, $p < 0.05$, but not overall performance, $Q_M[1] = 1.41$, $p = 0.24$, or auditory comprehension, $Q_M[1] = 0.23$, $p = 0.63$.

Premorbid Language Proficiency

Given that there were only four effect sizes in the higher L2 proficiency group, they were excluded from the analysis and described separately. In the whole group, premorbid language proficiency did not moderate either overall performance $Q_M[1] = 2.87$, $p = 0.09$, or auditory comprehension, $Q_M[1] = 0.05$, $p = 0.82$. Proficiency significantly moderated oral production, $Q_M[1] = 6.13$, $p < 0.05$. Individuals with equal proficiency in both languages had a significantly smaller difference between L1 and L2 in oral production, $RR = 1.03$ [0.95, 1.10], $p = 0.49$, compared to those who were more proficient in L1, $RR = 1.23$ [1.08, 1.38], $p < 0.01$. Individuals with higher L2 proficiency performed significantly better in L2 overall, $RR = 0.70$ [0.57, 0.87], $p < 0.01$, and in oral production, $RR = 0.71$ [0.53, 0.97], $p < 0.05$, but not in auditory comprehension, $RR = 0.88$ [0.75, 1.03], $p = 0.12$.

Given that 81% ($n = 30$) of the cases in the early AoA subgroup had equal proficiency and other two proficiency groups included three (L1) and four (L2) cases each, the moderator analysis was not performed and only descriptive statistics are reported here. Individuals with higher L1 proficiency performed comparably in the two languages overall, $RR = 1.04$ [0.80, 1.35], $p = 0.77$, as well as in auditory comprehension, $RR = 1.04$ [0.86, 1.26], $p = 0.67$, and oral production, $RR = 0.96$ [0.64, 1.44], $p = 0.84$. Similarly, participants with equal proficiency performed comparably overall, $RR = 1.03$ [0.95, 1.13], $p = 0.47$, as well as in auditory comprehension, $RR = 1.06$ [0.99, 1.15], $p = 0.10$, and oral production, $RR = 1.01$ [0.90, 1.13], $p = 0.84$. In contrast, the four participants with higher L2 proficiency described above were all early bilinguals; again, they showed significantly better performance in L2 overall, $RR = 0.70$ [0.55, 0.88], $p < 0.01$, as well as in oral production, $RR = 0.71$ [0.51, 0.99], $p < 0.05$, but not in auditory comprehension, $RR = 0.88$ [0.74, 1.04], $p = 0.13$.

In the late AoA subgroup, where 58% ($n = 33$) had equal proficiency and the rest ($n = 24$) had higher L1 proficiency, language proficiency did not moderate overall performance, $Q_M[1] = 1.60$, $p = 0.21$, and auditory comprehension, $Q_M[1] = 0.11$, $p = 0.74$, but was a significant moderator for oral production, $Q_M[1] = 5.49$, $p < 0.05$. Individuals with higher L1 proficiency performed significantly better in L1, $RR = 1.26$ [1.12, 1.42], $p < 0.001$, whereas the equal proficiency group showed comparable performance in oral production, $RR = 1.05$ [0.95, 1.16], $p = 0.36$.

Language Use

For the whole group, language use moderated overall performance, $Q_M[2] = 12.48$, $p < 0.01$, auditory comprehension, $Q_M[2] = 6.49$, $p < 0.05$, and oral production, $Q_M[2] = 6.29$, $p < 0.05$. Individuals who premorbidly used L1 more frequently had significantly greater magnitude of L1 advantage in overall performance, $RR = 1.19$ [1.09, 1.30], $p < 0.001$, as well as individuals who equally used both languages, $RR = 1.09$ [1.02, 1.16], $p < 0.05$, compared to the group with more L2 use who showed comparable performance in both languages, $RR = 0.95$

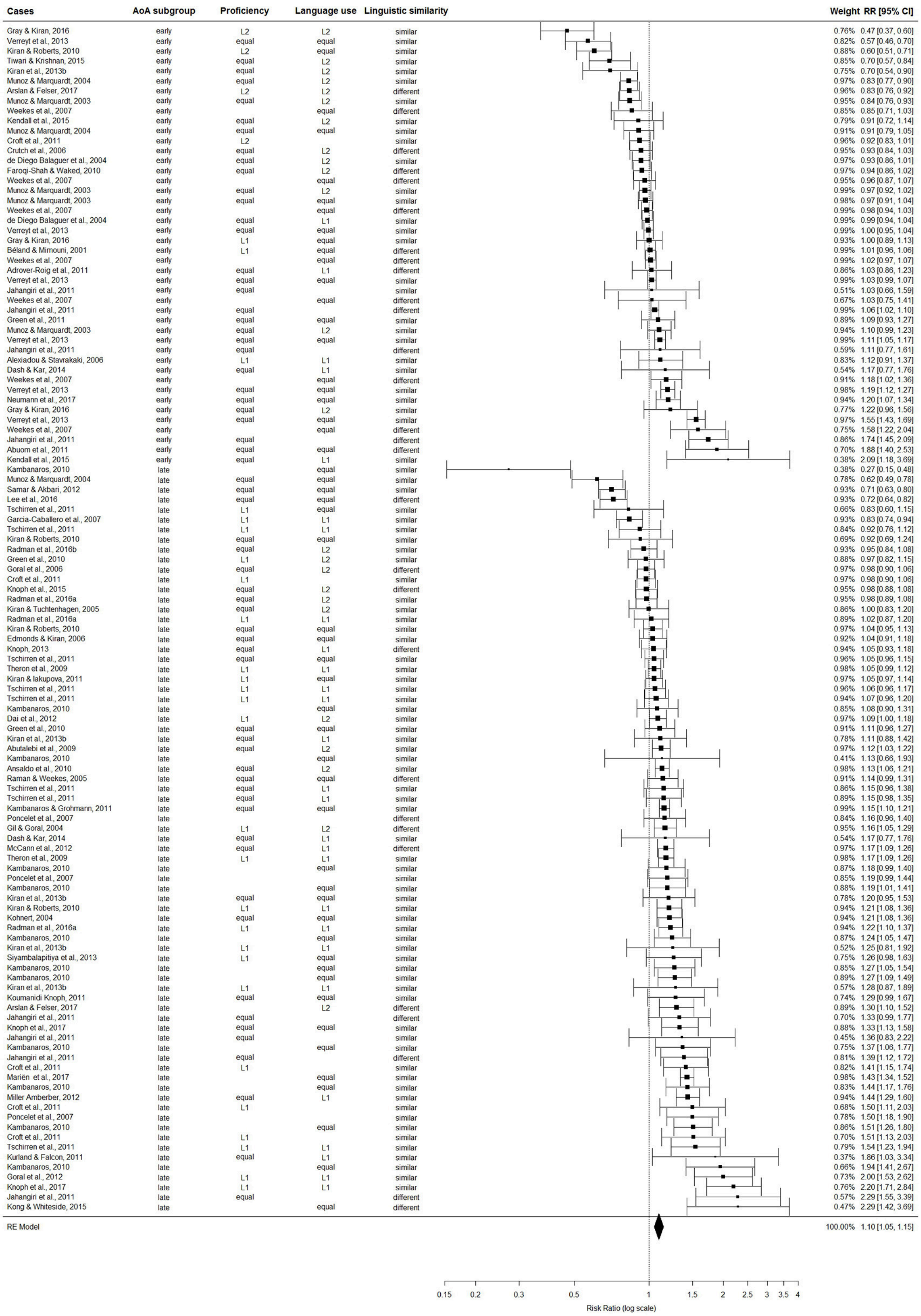


FIGURE 3 | For the whole trimmed sample ($k = 119$), the figure displays effect sizes (Risk Ratios) and corresponding 95% confidence intervals (CI) for the comparison between overall language performance in the earlier-acquired (L1) and later-learned (L2) languages. Values larger than one indicate better performance in L1 compared to L2 and values smaller than one indicate worse performance in L1 compared to L2.

TABLE 4 | Details of the moderator analysis.

Moderators	Overall performance						Auditory comprehension						Oral production					
	Effect size (RR)	95% CI	*k*	Q_M	*df*	*p*	Effect size (RR)	95% CI	*k*	Q_M	*df*	*p*	Effect size (RR)	95% CI	*k*	Q_M	*df*	*p*
WHOLE GROUP																		
AoA in years	**1.01****	[1.00, 1.01]	116	**8.84**	**1**	**<0.01**	1.00	[1.00, 1.01]	91	1.65	1	0.20	**1.01*****	[1.01, 1.02]	88	**13.61**	**1**	**<0.001**
AoA status			119	**11.37**	**1**	**<0.001**			91	2.59	1	0.11			91	**8.85**	**1**	**<0.01**
Early (AoA < 7 year)	1.00	[0.93, 1.07]	44				1.03	[0.97, 1.09]	38				0.97	[0.88, 1.07]	32			
Late (AoA ≥ 7 year)	**1.16*****	[1.10, 1.23]	75				**1.09*****	[1.04, 1.15]	53				**1.17*****	[1.09, 1.26]	59			
Premorbid proficiency			90	2.87	1	0.09			78	0.05	1	0.82			71	**6.13**	**1**	**<0.05**
Higher in L1	**1.16*****	[1.06, 1.26]	27				1.07	[0.99, 1.15]	25				**1.23****	[1.08, 1.38]	19			
Equal	**1.06***	[1.01, 1.12]	63				**1.08****	[1.03, 1.13]	53				**1.03**	[0.95, 1.10]	52			
Higher in L2	**0.70****	[0.57, 0.87]	4				0.88	[0.75, 1.03]	4				**0.71***	[0.53, 0.97]	3			
Language use			103	**12.48**	**2**	**<0.01**			79	**6.49**	**2**	**<0.05**			79	**6.29**	**2**	**<0.05**
More in L1	**1.19*****	[1.09, 1.30]	27				**1.11****	[1.03, 1.19]	22				**1.26****	[1.09, 1.46]	17			
Equal	**1.09***	[1.02, 1.16]	52				1.05	[0.99, 1.10]	34				1.07	[0.97, 1.17]	41			
More in L2	0.95	[0.87, 1.04]	24				0.98	[0.93, 1.04]	23				0.99	[0.87, 1.12]	21			
Linguistic similarity			119	0.35	1	0.55			91	0.45	1	0.50			91	0.06	1	0.81
Different	**1.12***	[1.03, 1.23]	29				**1.08***	[1.01, 1.16]	25				1.08	[0.94, 1.24]	18			
Similar	**1.09****	[1.03, 1.15]	90				**1.05***	[1.01, 1.10]	66				**1.10****	[1.03, 1.18]	73			
Age in years	**1.00****	[1.00, 1.01]	119	**8.71**	**1**	**<0.01**	**1.00***	[1.00, 1.01]	91	**5.70**	**1**	**<0.05**	1.00	[1.00, 1.01]	91	3.72	1	0.05
Years of education	**0.99***	[0.98, 1.00]	71	**3.90**	**1**	**<0.05**	1.00	[0.99, 1.01]	52	0.13	1	0.72	0.99	[0.97, 1.00]	63	3.04	1	0.08
Months post onsent	1.00	[1.00, 1.00]	106	0.97	1	0.32	1.00	[1.00, 1.00]	79	1.18	1	0.28	1.00	[1.00, 1.01]	85	2.04	1	0.15
EARLY AoA SUBGROUP																		
AoA in years	0.96	[0.91, 1.02]	44	1.72	1	0.19	0.97	[0.93, 1.01]	38	2.83	1	0.09	0.99	[0.91, 1.08]	32	0.06	1	0.80
Premorbid proficiency				-	-	-				-	-	-				-	-	-
Higher in L1	1.04	[0.80, 1.35]	3				1.04	[0.86, 1.26]	3				0.96	[0.64, 1.44]	2			
Equal	1.03	[0.95, 1.13]	30				1.06	[0.99, 1.15]	24				1.01	[0.90, 1.13]	27			
Higher in L2	**0.70****	[0.55, 0.88]	4				0.88	[0.74, 1.04]	4				**0.71***	[0.51, 0.99]	3			
Language use			34	**5.14**	**1**	**<0.05**			31	**6.57**	**1**	**<0.05**			24	0.50	1	0.48
More in L1	1.13	[0.90, 1.42]	5				0.97	[0.83, 1.13]	3				1.21	[0.86, 1.69]	4			
Equal	1.04	[0.94, 1.15]	21				1.06	[1.00, 1.12]	19				0.97	[0.82, 1.15]	14			
More in L2	**0.86***	[0.76, 0.98]	13				0.94	[0.88, 1.00]	12				0.88	[0.73, 1.07]	10			
Linguistic similarity			44	2.64	1	0.10			38	0.04	1	0.83			32	1.73	1	0.19
Different	1.08	[0.96, 1.22]	16				1.04	[0.95, 1.13]	15				1.10	[0.89, 1.34]	8			
Similar	0.96	[0.87, 1.05]	28				1.02	[0.95, 1.10]	23				0.94	[0.83, 1.06]	24			
Age in years	1.00	[1.00, 1.01]	44	1.20	1	0.27	1.00	[1.00, 1.01]	38	0.98	1	0.32	1.00	[0.99, 1.01]	32	0.09	1	0.76
Years of education	0.99	[0.97, 1.01]	26	0.74	1	0.39	1.00	[0.98, 1.01]	22	0.02	1	0.89	0.98	[0.95, 1.02]	19	0.60	1	0.44
Months post onsent	1.00	[1.00, 1.01]	32	1.31	1	0.25	1.00	[0.99, 1.00]	27	0.71	1	0.40	1.01	[1.00, 1.01]	27	3.15	1	0.08

(Continued)

TABLE 4 | Continued

LATE AoA SUBGROUP																		
AoA in years	1.00	[1.00, 1.01]	72	1.41	1	0.24	1	[1.00, 1.01]	53	0.23	1	0.63	**1.01***	[1.00, 1.02]	56	**6.15**	**1**	**<0.05**
Premorbid proficiency			57	1.60	1	0.21			51	0.11	1	0.74			42	**5.49**	**1**	**<0.05**
Higher in L1	**1.17*** **	[1.08, 1.28]	24				1.07	[0.99, 1.16]	22				**1.26*** **	[1.12, 1.42]	17			
Equal	**1.09***	[1.02, 1.17]	33				**1.09***	[1.02, 1.16]	29				1.05	[0.95, 1.16]	25			
Higher in L2	-	-	0				-	-	0				-	-	0			
Language use			64	2.77	2	0.25			45	3.46	2	0.18			51	2.20	2	0.33
More in L1	**1.20*** **	[1.09, 1.32]	22				**1.13****	[1.05, 1.23]	19				**1.28****	[1.09, 1.49]	13			
Equal	**1.12****	[1.04, 1.21]	31				1.03	[0.94, 1.12]	15				**1.13***	[1.01, 1.26]	27			
More in L2	1.06	[0.93, 1.19]	11				1.03	[0.95, 1.13]	11				1.10	[0.93, 1.30]	11			
Linguistic similarity			75	0.02	1	0.88			53	1.49	1	0.22			59	1.60	1	0.21
Different	**1.17***	[1.04, 1.32]	13				**1.16****	[1.04, 1.29]	10				1.07	[0.90, 1.26]	10			
Similar	**1.16*** **	[1.09, 1.23]	62				**1.07***	[1.02, 1.14]	43				**1.20*** **	[1.11, 1.29]	49			
Age in years	1.00	[1.00, 1.01]	75	3.36	1	0.07	1.00	[1.00, 1.01]	53	3.38	1	0.07	1.00	[1.00, 1.01]	59	2.40	1	0.12
Years of education	0.99	[0.98, 1.00]	45	2.36	1	0.12	0.99	[0.98, 1.01]	30	0.45	1	0.50	0.99	[0.98, 1.01]	44	1.39	1	0.24
Months post onset	1.00	[1.00, 1.00]	74	0.16	1	0.69	1.00	[1.00, 1.00]	52	0.79	1	0.38	1.00	[1.00, 1.00]	58	0.01	1	0.92

$^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Bold values represent statistically significant results.

[0.87, 1.04], $p = 0.25$. Individuals who premorbidly used L1 more frequently performed significantly better in L1 in both auditory comprehension, $RR = 1.11$ [1.03, 1.19], $p < 0.01$, and oral production, $RR = 1.26$ [1.09, 1.46], $p < 0.01$, compared to the L2 more frequent usage group who again showed comparable performance in both languages, (auditory comprehension: $RR = 0.98$ [0.93, 1.04], $p = 0.60$; oral production: $RR = 0.99$ [0.87, 1.12], $p = 0.84$). There were no significant differences between the more L1 and equal use groups for overall performance, auditory comprehension, and oral production.

Given that there were only five individuals in the early AoA group with greater L1 use, they were excluded from the moderator analysis and described separately. In the early AoA subgroup, language use moderated the outcomes for overall performance, $Q_M[1] = 5.14$, $p < 0.05$, and auditory comprehension, $Q_M[1] = 6.57$, $p < 0.05$, but not for oral production, $Q_M[1] = 0.50$, $p = 0.48$. In overall performance, individuals with more frequent L2 use showed significantly better performance in L2, $RR = 0.86$ [0.76, 0.98], $p < 0.05$, compared to those with equal use, who showed comparable overall performance, $RR = 1.04$ [0.94, 1.15], $p = 0.43$. Similarly for auditory comprehension, individuals with more L2 use showed comparable performance with a trend toward better performance in L2, $RR = 0.94$ [0.88, 1.00], $p = 0.07$, whereas the equal use group showed comparable performance with a trend toward better performance in L1, $RR = 1.06$ [1.00, 1.12], $p = 0.07$. The more L1 use group showed comparable performance overall, $RR = 1.13$ [0.90, 1.42], $p = 0.27$, in auditory comprehension, $RR = 0.97$ [0.83, 1.13], $p = 0.70$, and oral production, $RR = 1.21$ [0.86, 1.69], $p = 0.27$.

For the late AoA subgroup, language use did not moderate overall performance, $Q_M[2] = 2.77$, $p = 0.25$, auditory comprehension, $Q_M[2] = 3.46$, $p = 0.18$, or oral production, $Q_M[2] = 2.20$, $p = 0.33$.

Linguistic Similarity

For the whole group, binary linguistic similarity (different/similar languages) did not moderate overall performance, $Q_M[1] = 0.35$, $p = 0.55$, auditory comprehension, $Q_M[1] = 0.45$, $p = 0.50$, or oral production, $Q_M[1] = 0.06$, $p = 0.81$.

For the early AoA subgroup, binary linguistic similarity did not moderate overall performance, $Q_M[1] = 2.64$, $p = 0.10$, auditory comprehension, $Q_M[1] = 0.04$, $p = 0.83$, or oral production, $Q_M[1] = 1.73$, $p = 0.19$.

For the late AoA subgroup, binary linguistic similarity did not moderate overall performance, $Q_M[1] = 0.02$, $p = 0.88$, auditory comprehension, $Q_M[1] = 1.49$, $p = 0.22$, or oral production, $Q_M[1] = 1.60$, $p = 0.21$.

Similarly, linguistic similarity coded with three levels (very close/close/different languages) did not appear to be a significant moderator for overall performance (whole group: $k = 119$, $Q_M[2] = 0.78$, $p = 0.68$; early: $k = 44$, $Q_M[2] = 2.87$, $p = 0.24$; late: $k = 75$, $Q_M[2] = 1.96$, $p = 0.38$), auditory comprehension (whole group: $k = 91$, $Q_M[2] = 0.52$, $p = 0.77$; early: $k = 38$, $Q_M[2] = 0.05$, $p = 0.98$; late: $k = 53$, $Q_M[2] = 1.49$, $p = 0.48$), or oral production (whole group: $k = 91$, $Q_M[2] = 0.17$, $p = 0.92$;

early: $k = 32$, $Q_M[2] = 1.99$, $p = 0.37$; late: $k = 59$, $Q_M[2] = 2.08$, $p = 0.35$).

Additional Variables

In the whole sample, age moderated the outcomes for overall performance, $Q_M[1] = 8.71$, $p < 0.01$, and auditory comprehension, $Q_M[1] = 5.70$, $p < 0.05$: as age increased, the magnitude of L1 advantage increased. There was no significant moderation for oral production, $Q_M[1] = 3.72$, $p = 0.054$. Years of education moderated overall performance, $Q_M[1] = 3.90$, $p < 0.05$: as years of education increased, the magnitude of L1 advantage decreased. There were no significant effects of education either for auditory comprehension, $Q_M[1] = 0.13$, $p = 0.72$, or for oral production, $Q_M[1] = 3.04$, $p = 0.08$. Months post onset did not moderate overall performance, $Q_M[1] = 0.97$, $p = 0.32$, auditory comprehension, $Q_M[1] = 1.18$, $p = 0.28$, or oral production, $Q_M[1] = 2.04$, $p = 0.15$.

When the AoA subgroups were analyzed separately, age did not moderate outcomes for overall performance (early: $Q_M[1] = 1.20$, $p = 0.27$; late: $Q_M[1] = 3.36$, $p = 0.07$), auditory comprehension (early: $Q_M[1] = 0.98$, $p = 0.32$; late: $Q_M[1] = 3.38$, $p = 0.07$), or oral production (early: $Q_M[1] = 0.09$, $p = 0.76$; late: $Q_M[1] = 2.40$, $p = 0.12$). Similarly, years of education did not moderate overall performance (early: $Q_M[1] = 0.74$, $p = 0.39$; late: $Q_M[1] = 2.36$, $p = 0.12$), auditory comprehension (early: $Q_M[1] = 0.02$, $p = 0.89$; late: $Q_M[1] = 0.45$, $p = 0.50$), or oral production (early: $Q_M[1] = 0.60$, $p = 0.44$; late: $Q_M[1] = 1.39$, $p = 0.24$). Finally, months post onset did not moderate overall performance (early: $Q_M[1] = 1.31$, $p = 0.25$; late: $Q_M[1] = 0.16$, $p = 0.69$), auditory comprehension (early: $Q_M[1] = 0.71$, $p = 0.40$; late: $Q_M[1] = 0.79$, $p = 0.38$),or oral production (early: $Q_M[1] = 3.15$, $p = 0.08$; late: $Q_M[1] = 0.01$, $p = 0.92$).

DISCUSSION

The questions motivated this systematic review were whether people with aphasia are likely to exhibit better performance in their first-acquired (L1) than in a later-learned (L2) language, and whether age of acquisition (AoA), premorbid language proficiency, use and exposure, and linguistic similarity between the person's languages affect the consequences of aphasia in L1 and L2. We followed the PRISMA guidelines for a systematic review (Gates and March, 2016) and included 65 studies and 130 bilingual individuals with aphasia. Meta-analyses of effects sizes revealed the following answers to our questions.

L1 Primacy

We found that in the 119 bilingual speakers included in the analysis as a group, L1 was significantly better preserved than L2. This finding could be considered at odds with the view that different languages are processed in shared neural substrata for bilingual speakers (e.g., Abutalebi, 2008) and with the view held by many researchers and clinicians that bilingual people with aphasia tend to show equivalent language impairments after a stroke. The comparable impairment view has been supported by several reports in the literature. For example, Fabbro (2001) identified equivalent impairments in L1 and L2 in ≈60% of the cases he reviewed, who were early bilinguals with high proficiency in both languages. Unlike the findings reported by Fabbro (2001) and those reported in Albert and Obler (1978), our results appear to support Ribot (1882), which predicts that the earlier acquired language is more resistant to brain damage. This is also consistent with findings of better preservation in aphasia of words that are learned early in life compared to those learned later in life (for review see Brysbaert and Ellis, 2016).

We contend that our more rigorous analysis, which included a larger number of participants from a diversity of multilingual speakers of typologically different languages, is more reliable than the conclusions drawn from prior reviews. We note that the effect of language status (L1 vs. L2) confirmed here is often seen in case reports of bilingual speakers with aphasia but has rarely been analyzed according to the criteria developed in the present review.

Moreover, there has been a tendency in the literature on bilingual aphasia toward reporting performance of single cases according to the question of whether language impairments are parallel or differential (Paradis, 1983; Fabbro, 2001). We believe that posing this question can be misleading. It is critical to first determine whether parallel impairments should be expected, depending on the characteristics of the bilingual individuals. Indeed, it is possible that the reports of ≈60% of bilingual participants with aphasia demonstrating comparable impairments in both their L1 and L2 found in previous reviews are driven by early bilinguals and misrepresent the state of affairs for late bilinguals. We therefore divided our sample into early and late bilinguals to examine the observed difference between L1 and L2 separately for the two types of bilinguals. Furthermore, we examined whether additional bilingual characteristics, namely, specific AoA, frequency of language use, premorbid language proficiency, and linguistic similarity moderate the difference between L1 and L2.

Age of Language Acquisition

When we examined AoA as a binary categorical variable, our results demonstrated significant differences between early and late bilinguals. Specifically, late bilinguals, who acquired their other language after the age of seven, showed significantly better overall performance in L1 than in the later-learned language. In contrast, the early bilinguals who acquired their languages before the age of seven showed comparable performance in both languages. This result is consistent with previous findings from reports of balanced bilingual speakers who showed comparable levels of impairment (e.g., Fabbro, 2001; Kiran and Roberts, 2010). This difference between the two subgroups was significant despite the fact that the majority of individuals in both subgroups had equal pre-stroke proficiency in both their languages (81% and 58% in the early and late bilingual subgroups, respectively). Our finding of an effect of language status (i.e., significant difference between L1 and L2 performance) post stroke for late bilinguals challenges the assumptions of the shared bilingual neural substrate (SBNS) and the convergence hypothesis (Green, 2003). It is also at odds with the conclusions of Tschirren et al. (2011). It is possible that the differences in syntactic processing reported by Tschirren et al. (2011), together with generally comparable impairment, are the sort of outcomes that have contributed to the differential findings our meta-analysis revealed.

We found that, in the whole sample, AoA as a continuous variable moderated overall performance, oral production, but not auditory comprehension. This is consistent with findings that in bilinguals who are not highly proficient, language production is typically more difficult than language comprehension (e.g., Swain, 1985). It is possible that the substantial variance of performance among the late bilinguals (but not in the early bilinguals) included here allowed for the effect of AoA to emerge. Future studies could further examine the AoA at which the patterns of results change. Of interest, we found an effect of age, with older individuals showing the greater magnitude of L1 advantage compared to younger ones; the interaction of age and AoA could be further examined in future studies.

Thus, AoA moderated performance differences between L1 and L2 when early and late bilinguals were compared, which may suggest that a language that is acquired early enjoys a unique status and could potentially be differentially processed in the brain (e.g., Giussani et al., 2007). In contrast, the finding that AoA as a continuous variable significantly moderated only oral production in the late AoA subgroup only suggests that the exact AoA matters less. This is consistent with some views of the role of AoA in bilingualism (Birdsong and Molis, 2001). We note that we divided the participants into the early and late subgroups based on a theoretically motivated rationale. We found that in our sample, age 7 year was a natural breakpoint, considering that individuals started schooling in L2 at this age. A similar cut-off point (6 year) for early and later AoA was also used in the meta-analytic review on the bilingual advantage by Lehtonen et al. (2018).

Our findings of better overall performance in L1 than in L2 have implications for the cognitive neuropsychology of bilingual aphasia as well as for clinical aphasiology. Nevertheless, as expected, this finding was qualified by several variables identified in the literature as potential moderators: premorbid language proficiency, language use, and linguistic similarity (e.g., Goral et al., 2006; Ansaldo et al., 2008; Lorenzen and Murray, 2008). It can be argued that language proficiency and language use are typically correlated. As a rule, speakers who use a language with frequency and regularity are more likely to be highly proficient in that language (e.g., Gollan et al., 2015; Peñaloza et al., 2017). However, there are also instances in which people report greater use than proficiency, especially in L2. For the individuals included in the analysis in the current review, there was a significant association between these two variables ($n = 85$ $p < 0.01$, Cramer's $V = 0.34$); in our analyses, we examined the effects of language proficiency and language use separately.

Premorbid Language Proficiency

We tested whether premorbid language proficiency moderated the magnitude of the difference in performance between L1 and L2. One could assume that a premorbidly more proficient language would be better preserved after a stroke. Our results partially supported this hypothesis. Individuals with higher L1 proficiency and those with equal proficiency in their two languages showed the pattern observed for the sample as a whole, namely, better overall performance in L1 when compared to L2. There were only four individuals in the sample who reported higher premorbid L2 proficiency than L1 proficiency and they appeared to perform better in L2 compared to L1. No statistically significant differences were found between the higher premorbid L1 proficiency group compared with the equal proficiency group in overall performance and auditory comprehension scores, however the magnitude of L1 advantage in oral production scores was significantly greater for the group with higher L1 proficiency. These results overall suggest that L2 proficiency plays a role in the degree of impairment only when it surpasses the proficiency in L1. Given that the higher L2 proficiency group was very small in the present review, this assumption requires further investigation.

We also examined how proficiency moderated the effect of language status in the early bilingual and late bilingual subgroups separately. We observed that in the early bilingual group, individuals with higher L1 and with equal proficiency showed the pattern observed for the subgroup as a whole, namely, comparable performance in both languages. The four individuals who reported higher premorbid L2 proficiency were all early bilinguals and, as mentioned above, performed better in L2. There were no effects of proficiency in the late bilingual subgroup except for oral production, for which the magnitude of L1 advantage was significantly bigger for those individuals who reported higher L1 proficiency than those who reported to be equally proficient in both languages.

Thus, language proficiency appears to have a relatively small role in the results of overall differences between L1 and L2, except for those cases where L2 achieved higher proficiency than L1. This finding does not support the view that language proficiency has a greater role in determining language representation and processing in bilinguals than AoA (e.g., Perani et al., 1998; Abutalebi et al., 2001). We also found that the more years of education individuals had, the smaller was the magnitude of L1 advantage. This suggests that education in L2 could be used as an additional source of information for determining premorbid L2 proficiency.

It is of interest to note how language proficiency was measured in the reviewed studies. There was great variability in the measures and tools used (e.g., section A of the BAT; the Language Use Questionnaire, Muñoz et al., 1999), but generally, most studies included subjective self-ratings of the participants of their language abilities prior to the stroke. These self-ratings ranged in terms of the size of the scale and whether each ability was rated separately. In a few cases, family members' ratings were included as well. In none of the studies, formal measures of premorbid language abilities (e.g., language proficiency test, language placement test) were available.

Language Use

Language use has been discussed in recent publications on bilingual language performance (Linck et al., 2009), as a determining variable in degree of impairment as well as degree of recovery from aphasia (Goral et al., 2012, 2013; Knoph et al., 2017). This is particularly true for individuals who live in a monolingual L2 environment following immigration for example.

We examined whether the magnitude of the difference between L1 and L2 was influenced by language use. One could hypothesize that the more used language would be better preserved (Pitres, 1895/1983). Our results partially supported this hypothesis. In the whole group, those with more frequent use of L1 showed significantly better performance in L1 compared to L2 in all of the three outcomes, whereas those who rated their L2 use as more frequent than their L1 performed comparably in both languages in all three outcomes. For the early AoA subgroup, those who used L2 more often showed better performance in L2 based on overall performance scores, whereas those who used L1 more frequently and those who used both languages equally showed comparable performance in both languages in all three outcomes. Better performance in L2 compared to L1 was not found in the late bilinguals, whereas better performance in L1 and comparable performance were the typical patterns.

Similar to the findings for language proficiency in oral production performance, we found evidence of significantly greater magnitude of L1 advantage in the group with more frequent L1 use compared with the group where L2 was more frequently used, but not with the group where both languages were equally used. These findings suggest that language use affected the magnitude of L1 advantage when L2 became the most frequently used language. Thus, like premorbid language proficiency, language use has a moderating role on the findings, which does not seem to be independent of AoA.

Linguistic Similarity

There has been discussion in the literature regarding the degree to which language similarity influences the comparability of impairment in bilingual aphasia (Lorenzen and Murray, 2008). Whereas, on the one hand, one might predict that more similar languages would look similarly impaired following a stroke, there is little evidence to support this prediction and there is controversy in the literature regarding the role of language similarity on the neuronal organization of the languages of a bilingual (Kumar, 2014; Wong et al., 2016). On the other hand, one could assume that because linguistically similar languages share a significant portion of lexico-semantic representation (e.g., cognates), more cognitive control may be required to overcome cross-language interference. Our analyses revealed no effect of linguistic similarity. This finding is consistent with recent studies that attributed greater importance of language proficiency and use than of linguistic similarity (e.g., Muñoz and Marquardt, 2003; Ansaldo and Saidi, 2014; Kastenbaum et al., 2019). The finding is also consistent with neuroimaging studies that have demonstrated overlap in processing and representation among languages of bilinguals even for those who use languages that are very different from each other (e.g., Abutalebi et al., 2001; Wong et al., 2016).

We note, however, that quite a few studies have reported an effect of cognates, which is one aspect of language similarity that has been studied in aphasia (Kohnert, 2004; Kurland and Falcon, 2011; Kendall et al., 2015). Our finding of no role of linguistic similarity could be considered in opposition to such studies. It is possible that linguistic similarity affects the manifestation of specific linguistic aspects, consistent with findings that reported interference between languages that are similar (Fabbro, 2001; Goral et al., 2006), but that the degree of language similarity does not affect overall relative levels of impairment. Thus, it may be that effects of linguistic similarity on performance will be evident in tasks that require syntactic processing for languages that share or differ in specific morpho-syntactic aspects (e.g., Nilipour and Paradis, 1995; Yiu and Worrall, 1996; Goral et al., 2010) and in those that demand lexical-semantic processing for languages that share more or fewer cognates (Kohnert, 2004; Kurland and Falcon, 2011). We also note that dissociations in performance for bilingual patients with reading and writing disorders suggest that language type can constrain patterns of bilingual aphasia (see Weekes, 2012; Goral, 2019).

LIMITATIONS

The number of significant effect sizes we found points to the robustness of our findings, although the greater L2 proficiency results were based on a small number of cases and should be interpreted with caution. Furthermore, there was great variability among the studies included in the review, both in terms of the participants' characteristics and the language performance measures (see **Table 1**). Indeed, the variability of measures used is a limitation of the present data as well as of the field in general. Our data highlight the importance of greater uniformity of assessment in bilingual aphasia, which was one rationale for the development of the BAT (Paradis and Libben, 1987), although, other tools are clearly needed to assess specific languages and linguistic aspects.

We are mindful of drawing conclusions from the data beyond the domain of bilingual aphasia. Our results confirm the view that individual differences in the unique language background characteristics of a bilingual speaker are very likely to impact on the presentation of aphasia in more than one language. Additionally, an open question to date is to what degree differences in performance between L1 and L2 in late bilinguals are due to differential impairment levels or to differential pre-stroke mastery levels. Another conclusion which should be viewed with caution is the one regarding AoA. Although the transformation of the AoA into the binary classification (early/late) based on 7 year of age was motivated theoretically, other cut-off points can be considered in the future research. Furthermore, a lack of effects of specific AoA within the AoA subgroups could partially be a result of relatively low inter-individual variability in this variable.

Finally, the analyses we conducted did not allow us to consider in-depth language impairment patterns of multilingual individuals with aphasia, such as for instance, uncontrollable language blending and antagonistic recovery (Paradis, 1977, 2001), which are of great importance for understanding cognitive mechanisms of language. Moreover, given the cross-sectional nature of the present study, it does not inform us about the dynamics of language performance, which was an important aspect in the classification of recovery patterns in multilingual aphasia (Paradis, 1977, 2001).

CONCLUSION

To conclude, the current systematic review and meta-analysis revealed a better performance in L1 compared to L2 in bilingual speakers with aphasia. It also demonstrated that the magnitude of this difference was moderated by whether the bilinguals learned their two languages early in childhood or later. The better performance in L1 was a robust finding, which was moderated by premorbid language proficiency and frequency of use. Finally, linguistic similarity did not appear to interact with the magnitude of the difference in performance between L1 and L2.

The results we report here from a meta-analysis reflect the patterns observed in case studies, case-series, and group studies of multilingual individuals with aphasia. Cognitive neuropsychology has been a dominant theoretical movement in the study of aphasia for nearly 50 years. One defining feature of cognitive neuropsychology is the study of the single case and its bedrock assumption is that group studies are not meaningful because they average data across participants and consequently mask individual differences (Caramazza, 1986; though see Caplan, 1988; Grodzinsky et al., 1999). In the past decade, cognitive neuropsychologists have evolved toward advocating a case-series approach which retains the individual differences in single cases while accommodating the general patterns of performance in clinical groups (Schwartz and Dell, 2010; Rapp, 2011). For this reason, the problems of averaging that are debated extensively in the cognitive neuropsychological literature (e.g., McCloskey and Caramazza, 1988) do not apply to case-series designs. We contend that the type of meta-analysis conducted here also retains the individual patterns of performance.

Our findings reinforce the calls for (1) assessing all languages and collecting language background information (e.g., language use, premorbid language proficiency) of multilingual speakers with aphasia to obtain the most accurate assessment of their language abilities and (2) reporting performance in a way allowing researchers to compare the records among different studies, i.e., disclosing names of the assessment tools and scales used.

Growing understanding of the roles of such variables as premorbid language proficiency, language use and exposure, AoA, and structural similarities between one's languages will improve assessment practices and management options for multilingual speakers with aphasia. At the very least, multilingual speakers with aphasia should be assessed and treated with the understanding that it could be their earlier-acquired language that may be the key to greater success in restoring communication abilities.

DATA AVAILABILITY

The dataset and R script used for the analysis are provided in **Supplementary Material**.

AUTHOR CONTRIBUTIONS

EK contributed to this study by designing the study concept, searching the databases, screening the data, coding the studies, planning and conducting the data analysis, interpreting the results, and writing the manuscript. MG contributed by designing the study concept, coding the studies, interpreting the results, writing the manuscript, and providing the external assistance for preparing the manuscript. MN contributed to the study concept and coded the studies. BW contributed by inspiring the very idea of the study, designing the study concept, interpreting the results, and writing the manuscript.

FUNDING

The study was funded by the University of Oslo.

ACKNOWLEDGMENTS

Our sincere gratitude goes to Bård Uri Jensen and Katarina Antolovic for their help in preparing the present study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00445/full#supplementary-material

REFERENCES

*Abuom, T. O., Obler, L. K., and Bastiaanse, R. (2011). Using Swahili and English to test explanations of agrammatism. *Aphasiology* 25, 559–575. doi: 10.1080/02687038.2010.538417

Abutalebi, J. (2008). Language control and lexical competition in bilinguals: an event-related fMRI study. *Cereb. Cortex* 18, 1496–1505 doi: 10.1093/cercor/bhm182

Abutalebi, J., Cappa, S. F., and Perani, D. (2001). The bilingual brain as revealed by functional neuroimaging. *Bilingualism Lang. Cogn*. 4, 179–190. doi: 10.1017/S136672890100027X

Abutalebi, J., and Green, D. (2007). Bilingual language production: the neurocognition of language representation and control. *J. Neurolinguistics* 20, 242–275. doi: 10.1016/j.jneuroling.2006.10.003

Abutalebi, J., and Green, D. (2008). Control mechanisms in bilingual language production: neural evidence from language switching studies. *Lang. Cogn. Process*. 23, 557–582. doi: 10.1080/01690960801920602

*Abutalebi, J., Rosa, P. A., Tettamanti, M., Green, D. W., and Cappa, S. F. (2009). Bilingual aphasia and language control: a follow-up fMRI and intrinsic connectivity study. *Brain Lang*. 109, 141–156. doi: 10.1016/j.bandl.2009.03.003

*Adrover-Roig, D., Galparsoro-Izagirre, N., Marcotte, K., Ferré, P., Wilson, M. A., and Inés Ansaldo, A. (2011). Impaired L1 and executive control after left basal ganglia damage in a bilingual Basque–Spanish person with aphasia. *Clin. Linguist. Phon*. 25, 480–498. doi: 10.3109/02699206.2011.563338

Albert, M. L., and Obler, L. K. (1978). *The Bilingual Brain: Neuropsychological and Neurolinguistic Aspects of Bilingualism*. New York, NY: Academic Press.

*Alexiadou, A., and Stavrakaki, S. (2006). Clause structure and verb movement in a Greek-English speaking bilingual patient with Broca's aphasia: evidence from adverb placement. *Brain Lang*. 96, 207–220. doi: 10.1016/j.bandl.2005.04.006

*Altman, C., Goral, M., and Levy, E. S. (2012). Integrated narrative analysis in multilingual aphasia: the relationship among narrative structure, grammaticality, and fluency. *Aphasiology* 26, 1029–1052. doi: 10.1080/02687038.2012.686103

Ansaldo, A. I., Marcotte, K., Scherer, L., and Raboyeau, G. (2008). Language therapy and bilingual aphasia: clinical implications of psycholinguistic and neuroimaging research. *J. Neurolingu*. 21, 539–557. doi: 10.1016/j.jneuroling.2008.02.001

Ansaldo, A. I., and Saidi, L. G. (2014). Aphasia therapy in the age of globalization: cross-linguistic therapy effects in bilingual aphasia. *Behav. Neurol*. 2014:603085. doi: 10.1155/2014/603085

*Ansaldo, A. I., Saidi, L. G., and Ruiz, A. (2010). Model-driven intervention in bilingual aphasia: evidence from a case of pathological language mixing. *Aphasiology* 24, 309–324. doi: 10.1080/02687030902958423

*Arslan, S., and Felser, C. (2017). Comprehension of wh -questions in Turkish–German bilinguals with aphasia: a dual-case study. *Clin. Linguist. Phon*. 2017, 1–21. doi: 10.1080/02699206.2017.1416493

Bakhtiar, M., Jafary, R., and Weekes, B. S. (2017). Aphasia in Persian: implications for cognitive models of lexical processing. *J. Neuropsychol*. 11, 414–435. doi: 10.1111/jnp.12095

*Béland, R., and Mimouni, Z. (2001). Deep dyslexia in the two languages of an Arabic/French bilingual patient. *Cognition* 82, 77–126. doi: 10.1016/S0010-0277(01)00148-2

Birdsong, D. (2006). Age and second language acquisition and processing: a selective overview. *Lang. Learn*. 56, 9–49. doi: 10.1111/j.1467-9922.2006.00353.x

Birdsong, D., and Molis, M. (2001). On the evidence for maturational constraints in second-language acquisition. *J. Mem. Lang*. 44, 235–249. doi: 10.1006/jmla.2000.2750

Brysbaert, M., and Ellis, A. W. (2016). Aphasia and age of acquisition: are early-learned words more resilient? *Aphasiology* 30, 1240–1263. doi: 10.1080/02687038.2015.1106439

Bub, D. (2011). Facing the challenge of variation in neuropsychological populations: lessons from biology. *Cogn. Neuropsychol*. 28, 445–450. doi: 10.1080/02643294.2012.671766

Caplan, D. (1988). On the role of group studies in neuropsychological and pathopsychological research. *Cogn. Neuropsychol*. 5, 535–547. doi: 10.1080/02643298808253273

Caramazza, A. (1984). The logic of neuropsychological research and the problem of patient classification in aphasia. *Brain Lang*. 21, 9–20. doi: 10.1016/0093-934X(84)90032–4

Caramazza, A. (1986). On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance. *Brain Lang*. 5, 41–66. doi: 10.1016/0278-2626(86)90061–8

Caramazza, A., and McCloskey, M. (1988). The case for single-patient studies. *Cogn. Neuropsychol*. 5, 517–527. doi: 10.1080/02643298808253271

Chee, M. W., Caplan, D., Soon, C. S., Sriram, N., Tan, E. W., Thiel, T., et al. (1999). Processing of visually presented sentences inMandarin and English with fMRI. *Neuron* 23, 127–137. doi: 10.1016/S0896-6273(00)80759-X

Chee, M. W., Weekes, B., Lee, K. M., Soon, C. S., Schreiber, A., Hoon, J. J., et al. (2000). Overlap and dissociation of semantic processing of characters English words and pictures: evidence from fMRI. *Neuroimage* 12, 392–403 doi: 10.1006/nimg.2000.0631

Coltheart, M. (2017). The assumptions of cognitive neuropsychology: reflections on Caramazza (1984); Caramazza (1986). *Cogn. Neuropsychol*. 34, 397–402.doi: 10.1080/02643294.2017.1324950

Conner, P., Goral, M., Anema, I., Borodkin, K., Haendler, Y., Knoph, M., et al. (2018). The role of language proficiency and linguistic similarity in crosslinguistic treatment effects in aphasia. *J. Clin. Linguist. Phon*. 32, 739–757. doi: 10.1080/02699206.2018.1435723

Costa, A., Calabria, M.,Marne, P., Hernández, M., Juncadella, M., Gascón-Bayarri, J., et al. (2012). On the parallel deterioration of lexico-semantic processes in the bilinguals' two languages: evidence from Alzheimer's disease. *Neuropsychologia* 50, 740–753. doi: 10.1016/j.neuropsychologia.2012.01.008

*Croft, S., Marshall, J., Pring, T., and Hardwick, M. (2011). Therapy for naming difficulties in bilingual aphasia: which language benefits? *Int. J. Lang. Commun. Disord*. 46, 48–62. doi: 10.3109/13682822.2010.484845

*Crutch, S. J., Ridha, B. H., and Warrington, E. K. (2006). The different frameworks underlying abstract and concrete knowledge: evidence from a bilingual patient with a semantic refractory access dysphasia. *Neurocase* 12, 151–163. doi: 10.1080/13554790600598832

*Dai, E. Y. L., Kong, A. P. H., and Weekes, B. S. (2012). Recovery of naming and discourse production: a bilingual anomic case study. *Aphasiology* 26, 737–756. doi: 10.1080/02687038.2011.645013

*Dash, T., and Kar, B. R. (2014). Bilingual language control and general purpose cognitive control among individuals with bilingual aphasia: evidence based on negative priming and flanker tasks. *Behav. Neurol*. 2014:679706. doi: 10.1155/2014/679706

*de Diego Balaguer, R., Costa, A., Sebasti'an-Gall'es, N., Juncadella, M., and Caramazza, A. (2004). Regular and irregular morphology and its relationship with agrammatism: evidence from two Spanish-Catalan bilinguals. *Brain Lang*. 91, 212–222. doi: 10.1016/j.bandl.2004.02.007

*Edmonds, L. A., and Kiran, S. (2006). Effect of semantic naming treatment on crosslinguistic generalization in bilingual aphasia. *J. Speech Lang. Hear. Res*. 49:729. doi: 10.1044/1092-4388(2006/053)

Fabbro, F. (2001). The bilingual brain: bilingual aphasia. *Brain Lang*. 79, 201–210. doi: 10.1006/brln.2001.2480

Faroqi-Shah, Y., Frymark, T., Mullen, R., and Wang, B. (2010). Effect of treatment for bilingual individuals with aphasia: a systematic review of the evidence. *J. Neurolinguist*. 23, 319–341. doi: 10.1016/j.jneuroling.2010.01.002

*Faroqi-Shah, Y., and Waked, A. N. (2010). Grammatical category dissociation in multilingual aphasia. *Cogn. Neuropsychol*. 27, 181–203. doi: 10.1080/02643294.2010.509340

Fleiss, J., Levin, B., and Paik, M. (2003). *Statistical Methods for Rates and Proportions, 3rd Edn*. New York, NY:Wiley and Sons.

*García-Caballero, A., García-Lado, I., González-Hermida, J., Area, R., Recimil, M. J., Juncos Rabadán, O., et al. (2007). Paradoxical recovery in a bilingual patient with aphasia after right capsuloputaminal infarction. *J. Neurol. Neurosurg. Psychiatry* 78, 89–91. doi: 10.1136/jnnp.2006.095406

Gates, N. J., and March, E. G. (2016). A Neuropsychologist' s guide to undertaking a systematic review for publication: making the most of PRISMA guidelines. *Neuropsychol. Rev*. 26, 109–120. doi: 10.1007/s11065-016-9318-0

*Gil, M., and Goral, M. (2004). Nonparallel recovery in bilingual aphasia: effects of language choice, language proficiency, and treatment. *Int. J. Bilingualism* 8, 191–219. doi: 10.1177/13670069040080020501

Giussani, C., Roux, F. E., Lubrano, V., Gaini, S. M., and Bello, L. (2007). Review of language organisation in bilingual patients: what can we learn fromdirect brain mapping? *Acta Neurochir*. 149, 1109–1116. doi: 10.1007/s00701-007-1266-2

Gollan, T. H., and Kroll, J. F. (2001). "Lexical access in bilinguals," in *A Handbook of Cognitive Neuropsychology:What Deficits Reveal About the HumanMind*, ed. B. Rapp (New York, NY: Psychology Press), 321–345.

Gollan, T. H., Starr, J., and Ferreira, V. S. (2015). More than use it or lose it: the number-of-speakers effect on heritage language proficiency. *Psychonomic Bull. Rev*. 22, 147–155. doi: 10.3758/s13423-014-0649-7

Goral, M. (2012). "Cross-language treatment effects in multilingual aphasia," in *Aspects of Multilingual Aphasia*, eds M. R. Gitterman, M. Goral, and L. K. Obler (Bristol, UK: Multilingual Matters), 106–121.

Goral, M. (2019). "Acquired reading disorders in bilingualism," in *Handbook of the Neuroscience of Multilingualism*, ed J. W. Schwieter (Hoboken, NJ: J-Blackwell), 592–607.

*Goral, M., Levy, E. S., and Kastl, R. (2010). Cross-language treatment generalization: A case of trilingual aphasia. *Aphasiology* 24, 170–187. doi: 10.1080/02687030902958308

*Goral, M., Levy, E. S., Obler, L. K., and Cohen, E. (2006). Cross-language lexical connections in the mental lexicon: evidence from a case of trilingual aphasia. *Brain Lang*. 98, 235–247. doi: 10.1016/j.bandl.2006.05.004

*Goral, M., Naghibolhosseini, M., and Conner, P. S. (2013). Asymmetric inhibitory treatment effects in multilingual aphasia. *Cogn. Neuropsychol*. 30, 564–577. doi: 10.1080/02643294.2013.878692

*Goral, M., Rosas, J., Conner, P. S., Maul, K. K., and Obler, L. K. (2012). Effects of language proficiency and language of the environment on aphasia therapy in a multilingual. *J. Neurolinguist*. 25, 538–551.doi: 10.1016/J.JNEUROLING.2011.06.001

*Gray, T., and Kiran, S. (2013). A theoretical account of lexical and semantic naming deficits in bilingual aphasia. *J. Speech Lang. Hear. Res*. 56, 1314–1327. doi: 10.1044/1092-4388(2012/12-0091)

*Gray, T., and Kiran, S. (2016). The relationship between language control and cognitive control in bilingual aphasia. *Bilingualism Lang. Cogn*. 19, 433–452. doi: 10.1017/S1366728915000061

Green, D. (2003). "Neural basis of lexicon and grammar in L2 acquisition: the convergence hypothesis," in *The Interface Between Syntax and the Lexicon in Second Language Acquisition*, eds R. van Hout, A. Hulk, F. Kuiken, and R. Towell (Amsterdam: John Benjamins), 197–208.

*Green, D. W., Grogan, A., Crinion, J., Ali, N., Sutton, C., and Price, C. J. (2010). Language control and parallel recovery of language in individuals with aphasia. *Aphasiology* 24, 188–209. doi: 10.1080/02687030902958316

*Green, D. W., Ruffle, L., Grogan, A., Ali, N., Ramsden, S., Schofield, T., et al. (2011). Parallel recovery in a trilingual speaker: the use of the bilingual aphasia test as a diagnostic complement to the comprehensive aphasia test. *Clin. Linguist. Phon*. 25, 499–512. doi: 10.3109/02699206.2011.560990

Grodzinsky, Y., Piñango, M. M., Zurif, E., and Drai, D. (1999). The critical role of group studies in neuropsychology: comprehension regularities in Broca's aphasia. *Brain Lang*. 67, 134–147. doi: 10.1006/brln.1999.2050

Grosjean, F. (1982). *Life with Two Languages: An Introduction to Bilingualism*. Cambridge, MA: Harvard University Press.

Grosjean, F. (2013). "Bilingualism: a short introduction," in *The Psycholinguistics of Bilingualism*, eds F. Grosjean and P. Li. Sussex (Oxford: Wiley-Blackwell), 7–25.

*Jahangiri, N., Azarpazhooh, M. R., Ghaleh, M., and Seifhashemi, F. (2011). Thalamic and striatocapsular bilingual aphasia. *Asia Pac. J. Speech Lang. Hear*. 14, 187–196. doi: 10.1179/jslh.2011.14.4.187

Junqué, C., Vendrell, J., and Vendrell, P. (1995). "Differential impairments and specific phenomena in 50 Catalan-Spanish bilingual aphasic patients," in *Aspects of Bilingual Aphasia*, ed M. Paradis (New York, NY: Elsevier Science Ltd), 177–209.

*Kambanaros, M. (2010). Action and object naming versus verb and noun retrieval in connected speech: comparisons in late bilingual Greek–English anomic speakers. *Aphasiology* 24, 210–230. doi: 10.1080/02687030902958332

*Kambanaros, M. (2016). Verb and noun word retrieval in bilingual aphasia: a case study of language- and modality-specific levels of breakdown. *Int. J.Bilingual Educ. Bilingualism* 19, 169–184. doi: 10.1080/13670050.2015.1037717

*Kambanaros, M., and Grohmann, K. K. (2011). Profiling performance in L1 and L2 observed in Greek-English bilingual aphasia using the bilingual aphasia test: a case study from Cyprus. *Clin. Linguist. Phon*. 25, 513–529. doi: 10.3109/02699206.2011.563899

*Kambanaros, M., Messinis, L., and Anyfantis, E. (2012). Action and object word writing in a case of bilingual aphasia. *Behav. Neurol*. 25, 215–222. doi: 10.3233/BEN-2012-119006

Kastenbaum, J., Bedore, L., Peña, E., Sheng, L., Mavis, I., Sebastian-Vaytadden, R., et al. (2019). The influence of proficiency and language combination on bilingual lexical access. *Bilingualism* 22, 300–330. doi: 10.1017/S1366728918000366

*Kendall, D., Edmonds, L., Van Zyl, A., Odendaal, I., Stein, M., and Van derMerwe, A. (2015). What can speech production errors tell us about cross-linguistic processing in bilingual aphasia? Evidence from four English/Afrikaans bilingual individuals with aphasia. *S. Afr. J. Commun. Disord*. 62, 1–10. doi: 10.4102/sajcd.v62i1.111

*Kiran, S., Grasemann, U., Sandberg, C., and Miikkulainen, R. (2013a). A computational account of bilingual aphasia rehabilitation. *Bilingualism* 16, 325–342. doi: 10.1017/S1366728912000533

*Kiran, S., and Iakupova, R. (2011). Understanding the relationship between language proficiency, language impairment and rehabilitation: evidence from a case study. *Clin. Linguist. Phon*. 25, 565–583. doi: 10.3109/02699206.2011.566664

*Kiran, S., and Lebel, K. (2007). Crosslinguistic semantic and translation priming in normal bilingual individuals and bilingual aphasia. *Clin. Linguist. Phon*. 21, 277–303. doi: 10.1080/02699200701243873

Kiran, S., Pena, E., Bedore, L., and Sheng, L. (2010). "Evaluating the Relationship between category generation and language use and proficiency," in *Paper presented at the Donostia Workshop on Neurobilingualism* (San Sebastian).

*Kiran, S., and Roberts, P. M. (2010). Semantic feature analysis treatment in Spanish–English and French–English bilingual aphasia. *Aphasiology* 24, 231–261. doi: 10.1080/02687030902958365

*Kiran, S., Sandberg, C., Gray, T., Ascenso, E., and Kester, E. (2013b). Rehabilitation in bilingual aphasia: evidence for within- and between-language generalization. *Am. J. Speech Lang. Pathol*. 22:85. doi: 10.1044/1058-0360(2013/12-0085)

*Kiran, S., and Tuchtenhagen, J. (2005). Imageability effects in normal Spanish–English bilingual adults and in aphasia: evidence from naming to definition and semantic priming tasks. *Aphasiology* 19, 315–327. doi: 10.1080/02687030444000778

*Knoph, M. I. K. (2013). Language intervention in Arabic-English bilingual aphasia: a case study. *Aphasiology* 27, 1440–1458. doi: 10.1080/02687038.2013.832139.

*Knoph, M. I. N., Lind, M., and Simonsen, H. G. (2015). Semantic feature analysis targeting verbs in a quadrilingual speaker with aphasia. *Aphasiology* 29, 1473–1496. doi: 10.1080/02687038.2015.1049583

*Knoph, M. I. N., Simonsen, H. G., and Lind, M. (2017). Cross-linguistic transfer effects of verb-production therapy in two cases of multilingual aphasia. *Aphasiology* 31:2017. doi: 10.1080/02687038.2017.1358447

*Kohnert, K. (2004). Cognitive and cognate-based treatments for bilingual aphasia: a case study. *Brain Lang*. 91, 294–302. doi: 10.1016/j.bandl.2004.04.001

*Kong, A. P., and Whiteside, J. (2015). Early recovery of a multi-lingual speaker with aphasia using Cantonese and English. *Speech Lang. Hear*. 18, 133–139. doi: 10.1179/2050572814Y.0000000059

Köpke, B., Schmid, M. S., Keijzer, M., and Dostert, S. (eds.). (2007). *Language Attrition: Theoretical Perspectives*. Amsterdam; Philadelphia, PA: John Benjamins.

*Koumanidi Knoph, M. I. (2011). Language assessment of a Farsi-Norwegian bilingual speaker with aphasia. *Clin. Linguist. Phon*. 25, 530–539. doi: 10.3109/02699206.2011.563900

Kumar, U. (2014). Effect of orthography over neural regions in bilinguals: Aa view from neuroimaging. *Neurosci. Lett*. 580, 94–99. doi: 10.1016/j.neulet.2014.07.053

*Kurland, J., and Falcon, M. (2011). Effects of cognate status and language of therapy during intensive semantic naming treatment in a case of severe nonfluent bilingual aphasia. *Clin. Linguist. Phon*. 25, 584–600. doi: 10.3109/02699206.2011.565398

Lambon Ralph, M. A., Patterson, K., and Plaut, D. C. (2011). Finite case-series or infinite single-case studies? Comments on "Case-series investigations in cognitive neuropsychology" by Schwartz and Dell (2010). *Cogn. Neuropsychol*. 28, 466–474. doi: 10.1080/02643294.2012.671765

*Lee, B., Moon, H. I., Lim, S. H., Cho, H., Choi, H., and Pyun, S. B. (2016). Recovery of language function in Korean-Japanese crossed bilingual aphasia following right basal ganglia hemorrhage. *Neurocase* 22, 300–305. doi: 10.1080/13554794.2016.1141966

Lehtonen, M., Soveri, A., Laine, A., Järvenpää, J., de Bruin, A., and Antfolk, J. (2018). Is bilingualism associated with enhanced executive functioning in adults? A meta-analytic review. *Psychol. Bull*. 144, 394–425. doi: 10.1037/bul0000142

*Levy, E. S., Goral, M., Castelluccio De Diesbach, C., and Law, F. (2011). Stronger accent following a stroke: the case of a trilingual with aphasia. *Clin. Linguist. Phon*. 25, 815–830. doi: 10.3109/02699206.2011.570408

Libben, M. (2017). "Non-selective language activation and bilingualism as the default mental lexicon," in *Bilingualism: A Framework for Understanding the Mental Lexicon*, eds M. Libben, M. Goral, and G. Libben (John Benjamins), 103–122.

Linck, J. A., Kroll, J. F., and Sunderman, G. (2009). Losing access to the native language while immersed in a second language: evidence for the role of inhibition in second- language learning. *Psychol. Sci*. 20, 1507–1515. doi: 10.1111/j.1467-9280.2009.02480.x

Lorenzen, B., and Murray, L. L. (2008). Bilingual aphasia: a theoretical and clinical review. *Am. J. Speech Lang. Pathol*. 17, 299–317. doi: 10.1044/1058-0360(2008/026)

*Mariën, P., van Dun, K., Van Dormael, J., Vandenborre, D., Keulen, S., Manto, M., et al. (2017). Cerebellar induced differential polyglot aphasia: a neurolinguistics and fMRI study. *Brain Lang*. 175, 18–28. doi: 10.1016/j.bandl.2017.09.001

*McCann, C., Lee, T., Purdy, S. C., and Paulin, A. K. (2012). The use of the bilingual aphasia test with a bilingual Mandarin-New Zealand English speaker with aphasia. *J. Neurolinguist*. 25, 579–587. doi: 10.1016/j.jneuroling.2011.04.001

McCloskey, M., and Caramazza, A. (1988). Theory and methodology in cognitive neuropsychology: a response to our critics. *Cogn. Neuropsychol*. 5, 583–623, doi: 10.1080/02643298808253276

*Meinzer, M., Obleser, J., Flaisch, T., Eulitz, C., and Rockstroh, B. (2007). Recovery from aphasia as a function of language therapy in an early bilingual patient demonstrated by fMRI. *Neuropsychologia*, 45, 1247–1256. doi: 10.1016/j.neuropsychologia.2006.10.003

Menn, L., and Obler, L. K. (eds.). (1990). *Agrammatic Aphasia: A Cross-Language Narrative Sourcebook*. Amsterdam; Philadelphia, PA: John Benjamins Publishing, CO. 1369–1389.

*Miller Amberber, A. (2012). Language intervention in French-English bilingual aphasia: evidence of limited therapy transfer. *J. Neurolingu*. 25, 588–614. doi: 10.1016/j.jneuroling.2011.10.002

Miozzo, M., Costa, A., Hernández, M., and Rapp, B. (2010). Lexical processing in the bilingual brain: evidence from grammatical/morphological deficits. *Aphasiology* 24, 262–287. doi: 10.1080/02687030902958381

*Munarriz, A., Ezeizabarrena, M. J., and Gutierrez-Mangado, M. J. (2016). Differential and selective morpho-syntactic impairment in Spanish-Basque bilingual aphasia. *Bilingualism* 19, 810–833. doi: 10.1017/S136672891400042X

*Muñoz, M. L., and Marquardt, T. P. (2003). Picture naming and identification in bilingual speakers of Spanish and English with and without aphasia. *Aphasiology* 17, 1115–1132. doi: 10.1080/02687030344000427

Muñoz, M. L., and Marquardt, T. P. (2004). The influence of language context on lexical retrieval in the discourse of bilingual speakers with aphasia. *J. Multilingual Commun. Disord*. 2, 1–17. doi: 10.1080/14769670310001625345

Muñoz, M. L., Marquardt, T. P., and Copeland, G. (1999). A comparison of the code switching patterns of aphasic and neurologically normal bilingual speakers of English and Spanish. *Brain Lang*. 66, 249–274. doi: 10.1006/brln.1998.2021

Nadeau, S. E. (2019). Bilingual aphasia: explanations in population encoding. *J. Neurolingu*. 49, 117–143. doi: 10.1016/j.jneuroling.2018.10.002

*Neumann, Y., Walters, J., and Altman, C. (2017). Codeswitching and discourse markers in the narratives of a bilingual speaker with aphasia. *Aphasiology* 31, 221–240. doi: 10.1080/02687038.2016.1184222

Nickels, L., Howard, D., and Best, W. (2011).On the use of different methodologies in cognitive neuropsychology: drink deep and from several sources. *Cogn. Neuropsychol*. 28, 475–485. doi: 10.1080/02643294.2012.672406

Nilipour, R., and Paradis, M. (1995). "Breakdown of functional categories in three Farsi–English bilingual aphasic patients," in Aspects of Bilingual Aphasia, ed M. *Paradis (Oxford, UK: Pergamon Press)*, 123–138.

Paradis, M. (1977). "Bilingualism and aphasia," in Studies in Neurolinguistics, H. Whitaker and H. *A.Whitaker (New York, NY: Academic Press)*, 65–121.

Paradis, M. (ed.). (1983). *Readings on Aphasia in Bilinguals and Polyglots*. Montreal, QC: Marcel Didier.

Paradis, M. (2001). "Bilingual and polyglot aphasia," in *Language and Aphasia*, ed R. S. Berndt (Amsterdam: Elsevier Science), 69–91.

Paradis, M. (2004). *A Neurolinguistic Theory of Bilingualism*. Amsterdam; Philadelphia, PA: John Benjamins.

Paradis, M., and Libben, G. (1987). *The Assessment of Bilingual Aphasia*. Hillsdale, NJ: Lawrence Erlbaum.

Peñaloza, C., Rao, L., and Kiran, S. (2017). "The association between language use and proficiency and L2 performance in bilingual adults with aphasia. Front. Hum. Neurosci," in *Conference Abstract: Academy of Aphasia 55th Annual Meeting*. (Baltimore, Maryland).

Perani, D., and Abutalebi, J. (2005). The neural basis of first and second language processing. *Curr. Opin. Neurobiol*. 15, 202–206. doi: 10.1016/j.conb.2005.03.007

Perani, D., Paulesu, E., Galles, N. S., Duoux, E., De-haene, S., Bettibardi, V., Cappa, S., et al. (1998). Bilingual brain: proficiency and age of acquisition in a second language. *Brain* 121, 1841–1852

Pitres, A. (1895/1983). "Aphasia in polyglots," in *Readings on Aphasia in Bilinguals and Polyglots*, ed M. Paradis (Montreal, QC: Marcel Didier), 26–49.

*Poncelet, M., Majerus, S., Raman, I., Warginaire, S., and Weekes, B. S. (2007). Naming actions and objects in bilingual aphasia: a multiple case study. *Brain Lang*. 103, 158–159. doi: 10.1016/j.bandl.2007.07.078

Pot, A., Keijzer, M., and De Bot, K. (2018). The language barrier in migrant aging. *Int. J. Bilingual Educ. Bilingual*. 2018, 1–19. doi: 10.1080/13670050.2018.1435627

*Radman, N., Mouthon, M., Di Pietro, M., Gaytanidis, C., Leemann, B., Abutalebi, J. et al. (2016a). The role of the cognitive control system in recovery from bilingual aphasia: a multiple single-case fMRI study. *Neural Plasticity* 2016:7086. doi: 10.1155/2016/8797086

*Radman, N., Spierer, L., Laganaro, M., Annoni, J. M., and Colombo, F. (2016b). Language specificity of lexical-phonological therapy in bilingual aphasia: a clinical and electrophysiological study. *Neuropsychol. Rehabil*. 26, 532–557. doi: 10.1080/09602011.2015.1047383

*Raman, I., and Weekes, B. S. (2005). Acquired dyslexia in a Turkish-English speaker. *Ann. Dyslexia* 55, 79–104. doi: 10.1007/s11881-005-0005-8

Rapp, B. (2011). Case-series in cognitive neuropsychology: promise, perils, and proper perspective. *Cogn. Neuropsychol*. 28, 435–444. doi: 10.1080/02643294.2012.697453

Ribot, T. A. (1882). *Diseases of Memory: An Essay in the Positive Psychology*. New York, NY: Appleton.

Roberts, P. M., and Deslauriers, L. (1999). Picture naming of cognate and non-cognate nouns in bilingual aphasia. *J. Commun. Disord*. 32, 1–23. doi: 10.1016/S0021-9924(98)00026-4

Roberts, P. M., and Kiran, S. (2007). "Assessment and treatment of bilingual aphasia and bilingual anomia," in *Ramos Speech and Language Disorders in Bilinguals*, eds A. Ardila and E. Ramos (New York, NY: Nova Science), 109–130.

Rochford, G., and Williams, M. (1965). Studies in the development and breakdown of the use of names. *J. Neurol. Neurosurg. Psychiatry* 28, 407–413.

Rothman, J. (2015). Linguistic and cognitive motivations for the Typological Primacy Model (TPM) of third language (L3) transfer: timing of acquisition and proficiency considered. *Bilingual. Lang. Cogn*. 18, 179–190. doi: 10.1017/S136672891300059X

*Samar, R. G., and Akbari, M. (2012). A language teacher in the haze of bilingual aphasia: a Kurdish-Persian case. *Proc. Soc. Behav. Sci*. 32, 252–257. doi: 10.1016/j.sbspro.2012.01.037

Schepens, J., Dijkstra, T., Grootjen, F., and van Heuven, W. J. (2013). Crosslanguage distributions of high frequency and phonetically similar cognates. *PLoS ONE* 8:e63006. doi: 10.1371/journal.pone.0063006

Schwartz, M. F., and Dell, G. S. (2010). Case-series investigations in cognitive neuropsychology. *Cogn. Neuropsychol*. 27, 477–494. doi: 10.1080/02643294.2011.574111

Shallice, T. (1979). Case study approach in neuropsychological research. *J. Clin. Neuropsychol*. 1, 183–211. doi: 10.1080/01688637908414450

*Siyambalapitiya, S., Chenery, H. J., and Copland, D. A. (2013). Lexicalsemantic representation in bilingual aphasia: findings from semantic priming and cognate repetition priming. *Aphasiology* 27, 1302–1321. doi: 10.1080/02687038.2013.817521

Swain, M. (1985). "Communicative competence: some roles of comprehensible input and comprehensible output in its development," in *Input in Second Language Acquisition*, eds S. Gass and C. Madden (Rowley, MA: Newbury House), 235–253.

*Theron, K., van der Merwe, A., Robin, D. A., and Groenewald, E. (2009). Temporal parameters of speech production in bilingual speakers with apraxic or phonemic paraphasic errors. *Aphasiology* 23, 557–583. doi: 10.1080/02687030701801717

*Tiwari, S., and Krishnan, G. (2015). Selective L2 cognate retrieval deficit in a bilingual person with aphasia: a case report. *Speech Lang. Hear*. 18, 243–248. doi: 10.1179/2050572815Y.0000000008

Tomoschuk, B., Ferreira, V. S., and Gollan, T. H. (2018). When a seven is not a seven: self-ratings of bilingual language proficiency differ between and within language populations. *Bilingualism* 2018, 1–21. doi: 10.1017/S1366728918000421

*Tschirren, M., Laganaro, M., Michel, P., Martory, M. D., Di Pietro, M., Abutalebi, J. et al. (2011). Language and syntactic impairment following stroke in late bilingual aphasics. *Brain Lang*. 119, 238–242. doi: 10.1016/j.bandl.2011.05.008

Ullman, M. T. (2001). A neurocognitive perspective on language: the declarative/procedural model. *Nat. Rev. Neurosci*. 2, 717–726. doi: 10.1038/35094573

*Verreyt, N., De Letter, M., Hemelsoet, D., Santens, P., and Duyck, W. (2013). Cognate effects and executive control in a patient with differential bilingual aphasia. *Appl. Neuropsychol.* 20, 221–230. doi: 10.1080/09084282.2012.753074

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* 36, 1–48. doi: 10.18637/jss.v036.i03

Vilariño, I., Prieto, J. M., Robles, A., Lema, M., and Noya, M. (1997). A study of bilingual Galician-Castilian aphasic patients. *Rev. Neurol.* 25, 1165–1167.

Weekes, B. S. (2005). Acquired disorders of reading and writing: crossscript comparisons. *Behav. Neurol.* 16, 51–57. doi: 10.1155/2005/492935

Weekes, B. S. (2010). Issues in bilingual aphasia: an introduction. *Aphasiology* 24, 123–125. doi: 10.1080/02687030902958274

Weekes, B. S. (2012). Acquired dyslexia and dysgraphia across scripts. *Behav. Neurol.* 25, 159–163. doi: 10.3233/BEN-2012-119000

*Weekes, B. S., Su, I.-F., Yin, W., and Zhang, X. (2007). Oral reading in bilingual aphasia: evidence from Mongolian and Chinese. *Bilingual. Lang. Cogn.* 10, 201–210. doi: 10.1017/s1366728907002945

*References marked with an asterisk indicate studies included in the present meta-analysis.

Wong, B., Yin, B., and O'Brien, B. (2016). Neurolinguistics: structure, function, and connectivity in the bilingual brain. *BioMed. Res. Int.* 2016, 1–22. doi: 10.1155/2016/7069274

Yiu, E. M. L., and Worrall, L. E. (1996). Agrammatic production: a crosslinguistic comparison of English and Cantonese. *Aphasiology* 10, 623–647. doi: 10.1080/02687039608248440

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Communication

ORIGINAL RESEARCH
published: 24 April 2019
doi: 10.3389/fcomm.2019.00014

Check for updates

Cognitive Reserve and Its Effect in Older Adults on Retrieval of Proper Names, Logo Names and Common Nouns

Sonia Montemurro[1,2], *Sara Mondini*[1,2]*, *Chiara Crovace*[1] *and Gonia Jarema*[3]

[1] Department of General Psychology, University of Padova, Padova, Italy, [2] Human Inspired Technology Research-Centre, University of Padova, Padova, Italy, [3] Université de Montréal and Research Centre, Institut Universitaire de Gériatrie de Montréal, Montréal, QC, Canada

OPEN ACCESS

Edited by:
Stefano F. Cappa,
Istituto Universitario di Studi Superiori di Pavia (IUSS), Italy

Reviewed by:
Maxime Montembeault,
University of California at San Francisco, United States
Stephanie Mathey,
Université de Bordeaux, France

***Correspondence:**
Sara Mondini
sara.mondini@unipd.it

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: 19 September 2018
Accepted: 28 March 2019
Published: 24 April 2019

Citation:
Montemurro S, Mondini S, Crovace C and Jarema G (2019) Cognitive Reserve and Its Effect in Older Adults on Retrieval of Proper Names, Logo Names and Common Nouns. Front. Commun. 4:14. doi: 10.3389/fcomm.2019.00014

Previous studies showed that high Cognitive Reserve (CR, years of education and experience and knowledge acquired in life) is correlated with language proficiency as measured with vocabulary size, verbal analogy, and semantic processing. The aim of the present study is to investigate the relationship between CR and the ability in retrieving different categories of words: Proper Names, Logo Names, and Common Nouns. The hypothesis is that CR contributes more in retrieving Common Nouns and Logo Names which are highly semantically interconnected, than retrieving Proper Names which are pure referring expressions. Forty-six Italian healthy older adults underwent the Montreal Cognitive Assessment (MoCA) and their performances spanned from low to high global cognitive profile. They were also administered a picture naming task for Proper Names, Logo Names and Common Nouns. Latency and Accuracy were recorded. CR was measured with the Cognitive Reserve Index (CRI) questionnaire which provides a measure of education, working time activities, and leisure time activities. Participants were significantly faster and more accurate in name retrieval when CR was high. CRI and MoCA as interaction terms predicted naming Latency with a stronger effect of CRI when the global cognitive profile was in the low range. The effect of CRI on Accuracy was lower for Proper Names than for Common Nouns and Logo Names, which did not differ from each other. Our results show that name retrieval Accuracy can be predicted by CR, significantly more in the case of Logo Names and Common Nouns than in the case of Proper Names. As Proper Names have scarce semantic interconnections and are arbitrarily assigned to unique individuals, they are not much influenced by CR. Although Logo Names are also arbitrarily assigned to their bearers, they can be conceptually categorized and thus influenced by reserve. The weak relationship between Proper Names and CR might suggest a proper name task as a useful tool to detect early signs of dementia, in particular for persons with high CR.

Keywords: proper names, common nouns, logo names, cognitive reserve, aging, naming

INTRODUCTION

Exposure to education, working activities and leisure time activities converge into a broader concept called "Cognitive Reserve" (CR) with a protective effect on cognitive functioning. CR mediates between a brain pathology and its clinical outcomes (Stern, 2006, 2012), so that people with high CR may achieve the ability to cope with age-related brain changes, thus delaying the symptoms of dementia (Scarmeas et al., 2001). In the case of globally healthy conditions, CR can provide flexibility in network selection, as it is characterized by networks developing over the lifespan as the result of innate processes (Barulli and Stern, 2013). Previous studies have shown that when the task demand increases, healthy adults recruit more brain networks (Ansado et al., 2013); thus, the load of a cognitive task is as relevant as the integrity of the global cognitive status. The relationship between CR and different types of tasks has shown that older adults with higher CR have a better performance in memory (Erber and Szuchman, 1996), in executive functioning tasks (Roldán-Tapia et al., 2012), as also in language tasks (Hultsch et al., 1993), compared to older adults with lower CR.

We want to investigate whether the CR of older adults and their global cognitive profile could act as predictors of name retrieval performance. Previous studies have already investigated name retrieval in aging, showing significant age-related effects (Rastle and Burke, 1996; Almond and Morrison, 2017). In this context, Almond and Morrison (2017) compared young and older adults in two experimental tests, involving the retrieval of proper names: (1) a face-name association task and (2) a pure-list task. Their results showed evidence of age-related deficits in the face-name association task, which however was claimed to be not highly sensitive for assessing age-related name recall deficits. In their study, however, names and faces used were not famous, either familiar (i.e., participants were instructed to learn names associated to new faces). For assessing name retrieval of familiar faces, some other researches have considered very famous people from local and international cultural settings (e.g., political figures, famous actors, religious figures, etc.), widely known outside their specific domain of fame (e.g., Semenza et al., 2003; Montemurro et al., 2018). In fact, persons with name retrieval deficits can find difficult to retrieve names of entities they know since long time.

Failure in name retrieval is one of patients' main subjective complaints, especially in the case of proper names (Cohen and Burke, 1993). Although deficits in picture naming task can be expected in the later stages of aging (see Feyereisen, 1997), cases of earlier name retrieval deficit are often reported, especially for the category of proper names (Greene and Hodges, 1996; Semenza et al., 2003). Whether CR can modulate proper name retrieval is a relatively recent question (Mondini and Semenza, 2016; Montemurro et al., 2018), which we want to address considering also the global cognitive conditions.

Another research question is to evaluate whether different cognitive processes are diversely sensitive to CR. In fact, it seems that the beneficial effects of high CR do not necessarily help performance in all cognitive tasks; for example, math performance has been shown to be independent from the level of CR in older adults from 65 years old (Arcara et al., 2017).

In order to make a timely diagnosis of dementia, it is important to identify early signs of decline also in patients whose symptoms may be masked by high-level CR (Robertson, 2013). Semenza et al. (2003) have shown that one of the first signs of Alzheimer's dementia can be proper name anomia. However, Mondini and Semenza (2016) and Montemurro et al. (2018) have later reported that proper name retrieval is hardly related to CR.

Mondini and Semenza (2016) analyzed the performance of 40 mildly cognitive impairment patients and first showed that, whereas CR was positively correlated to better global cognitive profile as assessed by the MMSE, CR did not predict name retrieval of famous people in a paper pencil task (Semenza et al., 2003). Authors interpreted this finding as due to the arbitrary link between proper names and bearers. Montemurro et al. (2018) further investigated this phenomenon with a more controlled experimental setting. They used the Italian version (Conti et al., 2015) of the Montreal Cognitive Assessment (MoCa, Nasreddine et al., 2005), which is a more sensitive measure of global cognitive profile, in two groups of participants: healthy elderly and patients with dementia. The correlation between CR and global cognitive profile was confirmed (in patients), while CR did not play any role in naming famous faces in both patients and healthy elderly. This restated the hypothesis of a weak effect of CR in naming proper names due to their poor interconnections in the semantic system, although a preserved cognitive profile. However, in both these studies, only the proper name category was tested, and both used a paper-and-pencil task (Semenza et al., 2003). The present study includes, instead, some novelties: the experimental paradigm is computer-based which is more precise and suitable for stimulus control and repeated measure analysis (Maruish and Moses, 1996). Furthermore, we included a set of Common Nouns and the captivating category of Logo Names.

Darby et al. (2017) showed that patients with Alzheimer's dementia and patients with mild cognitive impairment could benefit from CR in those tasks that require the involvement of executive and semantic functions. The semantic requirement might explain the weak relationship between CR and proper names (Montemurro et al., 2018). Indeed, semantic operations necessary for naming proper names may, at some point of the retrieval process, be different from naming other categories of names. In contrast to common nouns, proper names denote individual entities, and the set of attributes labeled by a proper name are related to one another only by virtue of belonging to unique entities (Semenza, 2009). In this context, different ways of possessing reference might be reflected in different mechanisms of semantic memory: "individual semantics" may refer to proper names, whereas "general semantics" to common nouns (Semenza, 2009).

The theoretical distinction between proper names and common nouns has a long history (e.g., Searle, 1969; Kripke, 1980). These philosophers described the linguistic properties of proper names defining them as expressions which convey reference but not sense. Later, a series of experimental studies contributed to this issue (e.g., Gorno-Tempini et al., 1998; Evrard, 2002; Pelamatti et al., 2003; Semenza, 2006, 2009), reporting disproportionate age-related problem in lexical access to proper names compared to common nouns, especially in production (Brédart et al., 1997). For example, the study of Evrard (2002) showed that healthy elderly compared to younger adults may experience more tip-of-the-tongue states for proper names than for common names. In contrast to common nouns, whose attributes are linked with each other in rich semantic interactions, the link between proper names and their bearers has been considered weak (e.g., Semenza and Zettin, 1989; Burke et al., 1991; Semenza, 2009).

It is well known that proper names possess particular linguistic features, as do brand names (Gontijo et al., 2002), which can also be very familiar, due to people's exposure to commerce and advertising communication. Logo names, a type of brand name, are arbitrarily assigned to products, companies or associations and sometimes acquire popularity. They do not designate unique entities and can be conceptually categorized, similarly to common nouns (Gontijo et al., 2002). Both proper names and some logo names can be considered, at the level of retrieval, pure referential expressions; however, logo names (similarly to common nouns) are generally more related to semantic facts (i.e., repeated and shared knowledge) rather than to episodic facts (i.e., facts related to individual episodic experience with the bearer). In other words, some logo names become very common in our social environment so that they can be conceptually categorized (Gontijo et al., 2002); for example, saying "Mercedes" while looking at its logo, may recall not only car-specific features, but also various models of the same car, which may have different engine powers, different colors, etc.

In the present study we examine, in terms of Latency and Accuracy, the effect of CR on name retrieval in a group of healthy older adults. Assuming a differential semantic interconnection between nouns and bearers, our hypothesis is that, as it is related to semantic richness, CR could show a differential influence in retrieving names with different degrees of semantic name-bearer interconnection. Thus, three categories of names were used in this study: (1) Proper Names, (2) Logo Names, and (3) Common Nouns. We predict that the higher the semantic interconnection between target and bearer, the higher the influence of CR on the task. Logo Names and Common Nouns, which are better related to repeated and shared knowledge of their bearers, are predicted to be more influenced by CR than Proper Names, which are instead weakly connected with their bearers.

METHODS

Participants

A total of forty-six Italian native speakers (28 women, 18 men) aged from 65 to 96 years and with 3 to 21 years of education participated in the study (**Table 1**).

An informal interview allowed to record information about their medical history, which showed no symptoms of psychiatric disease or neurological impairment. They were administered the MoCA, the Cognitive Reserve Index questionnaire (CRIq, Nucci et al., 2012) and a Picture Naming task.

TABLE 1 | Descriptive data of participants: Age, years; Education, years of formal education; MoCA, raw scores at the MoCA test (Nasreddine et al., 2005); CRI, Cognitive Reserve Index (from the Cognitive Reserve Index questionnaire; Nucci et al., 2012); M, mean; SD, standard deviation.

	Participants (N = 46)		
	M	**SD**	**Range**
Age	81.09	7.73	65–96
Education	8.89	4.61	3–21
MoCA	20.63	4.01	14–27
CRI	97.91	23.63	59–152

Materials and Procedure

The MoCA test is a brief neuropsychological tool, which provides a global cognitive profile. It consists of eight sub-tests tapping different cognitive domains (i.e., memory, language, visuospatial skills, executive functions, and orientation in time and space). The MoCA is widely used in clinical practice and is very sensitive to mild cognitive impairment in aging, especially in neurodegenerative diseases. Its administration lasts about 10 min and the maximum score is 30. As reported in **Table 1**, the range of participants' raw scores on the MoCA test ranged from 14 to 27. Each raw score was then adjusted for age and education according to the Italian normative data of Conti et al. (2015). All participants' scores fell within the non-clinical population (i.e., above the adjusted Italian cut-off of 17.36). More specifically, considering the global cognitive performance of our participants, the 19.56% of their adjusted scores fell within the borderline/fragile (but not pathological) group, and the rest of the participants fell within the preserved elderly.

Cognitive Reserve was measured with the Cognitive Reserve Index questionnaire (Nucci et al., 2012), which is a semi-structured interview. It requires approximately 10 min to complete and includes 20 questions grouped into three sections: Education (*CRI-Education*), Working activities (*CRI-WorkingActivity*), and Leisure time activities (*CRI-LeisureTime*). *CRI-Education* is made up of years of formal education and any additional training courses. *CRI-WorkingActivity* refers to the cognitive load and personal responsibility of an occupation, combined with the number of years it has been carried out. Finally, *CRI-LeisureTime* measures the frequency and the amount of intellectual, social and physical activities (e.g., reading newspapers or books, playing music, participation in charitable activities, traveling, doing sports, etc.). Additionally, the questionnaire includes items about life-long experiences that require a certain cognitive load (e.g., years of bank account management). The total CRIq score is an estimation of Cognitive Reserve. It is the average of the three sub-scores standardized and transposed to a scale with a mean of 100 and a standard deviation of 15 (Nucci et al., 2012). This standardized index of Cognitive Reserve (i.e., CRI) is informative in both clinical and research contexts; it derives from the combination of exposures to life activities over time (see Stern, 2006 for more details). Its total score is adjusted for age via a regression-based method to allow comparisons between groups of different ages (see Nucci et al., 2012 for further details).

The Picture Naming task was built to measure naming Latency and naming Accuracy. Picture selection was based on a total of 159 colored images, preliminarily rated on a 1–7 Likert scale according to Familiarity, Difficulty of naming and Age of Acquisition. Pictures represented three categories of entities: very famous persons, well-known Logos, and living, and non-living things.

Pictures of famous persons were chosen from a variety of settings, such as movies (e.g., "Sean Connery," "Marilyn Monroe"), politics (e.g., "Vladimir Putin," "Angela Merkel"), and religious contexts (e.g., "Pope Francis," "Mother Teresa"), to make sure that each of them was known by a large majority of people.

Logos were chosen from a wide range of international and local symbols whose visual representation was very frequent both in Italian and International settings (e.g., sport brands such as "Nike," "Adidas"), car companies (e.g., "Audi" and "Mercedes"), commercial products (e.g., "Rolex" and "Benetton"). We excluded Logos whose visual representation carries the meaning of the name (e.g., "Apple") and Logos whose visual representation carries the initial letter of the name (e.g., "McDonald's").

The pictures of living and non-living things for Common Nouns derived (with few adjustments) from a set of 360 high-quality color images (Moreno-Martínez and Montoro database, 2012).

A group of forty healthy Italian individuals (15 women, 25 men), with no history of neurological or psychiatric disease (mean age = 73.5 ± 7.6; mean education = 10.2 ± 5.1) rated all the pictures. Participants were asked to judge them on a 1–7 Likert scale according to Familiarity, Difficulty of naming and Age of Acquisition. For Familiarity, participants were asked to rate each picture from 1 to 7, where 1 indicated "completely unknown" and 7 indicated "highly familiar" (e.g., Valentine, 1998; Salmon et al., 2010). For Difficulty, participants were asked to rate each image from 1 to 7, where 1 indicated "impossible to name" and 7 indicated "very easy to name" pictures (e.g., Moreno-Martínez and Montoro, 2012). For Age of Acquisition, participants were asked to rate each image from 1 to 7, where 1 indicated "never acquired" and 7 indicated "acquired very early": before 3 years of age (e.g., Valentine, 1998; Salmon et al., 2010). Examples were provided before starting the rating phase. A 1–7 Likert scale was visually available to participants, for each variable, during the entire scoring process. The rating allowed to eventually select 30 pictures for each of the three categories (i.e., Proper Names, Logo Names, and Common Nouns) for the computerized Picture Naming task.

In the Common Noun picture category, Familiarity showed a mean of 5.89 (±0.98; range: 3.33–6.93); Difficulty a mean of 5.08 (± 0.79; range: 3.2 to 5.98); Age of Acquisition a mean of 3.97 (± 0.97; range: 2.2–5.73). In the Proper Name picture category, Familiarity showed a mean of 6.14 (± 0.58; range: 5.05 to 7.01; Difficulty a mean of 5.16 (± 0.54; range: 4.15–6.25); Age of Acquisition a mean of 2.08 (± 0.22; range: 1.75–2.65). In the Logo Name picture category, Familiarity showed a mean of 3.68 (± 1.59; range: 1.53–6.53); Difficulty, a mean of 2.89 (±1.41; range: 0.9–5.53); Age of Acquisition a mean of 1.68 (±0.44; range: 1.13–2.6).

As expected, Logo Name pictures were rated as less familiar than both Proper Names ($B = -2.47$, $p < 0.001$) and Common Noun pictures ($B = -2.47$, $p < 0.001$), which did not differ from each other ($B = 0.26$, $p = 0.36$). Logo Name images were the most difficult to retrieve compared to both famous faces ($B = 2.27$, $p < 0.001$) and Common object images ($B = 2.18$, $p < 0.001$), which did not differ from each other ($B = 0.09$, $p < 0.71$). Finally, Common Noun pictures were acquired significantly earlier than both famous face images ($B = -1.88$, $p < 0.001$) and Logo pictures ($B = -2.3$, $p < 0.001$), while famous face images were acquired earlier than Logo Name pictures ($B = -0.4$, $p < 0.01$).

To sum up, Logo pictures were less familiar, later acquired and more difficult to name than the other two categories. While pictures of famous faces and pictures of Common objects were balanced for the two first psycholinguistic variables, Age of Acquisition was, as expected, higher for famous faces.

The final set of the computerized Picture Naming task consisted of 90 colored images: 30 famous faces for Proper Names, balanced across international/local and male/female characters; 30 pictures for Common Nouns, balanced across living/non-living things; 30 pictures of logos for Logo Names, balanced across international/local subjects (see **Figure 1** for some examples).

Each participant attended a training session before performing the experimental session. Instructions were repeated if necessary. Image dimension was standardized with GIMP software in a 400 × 400 pixel-frame. Pictures were presented randomly; a fixation point appeared on the center of the computer screen for 500 milliseconds, followed by a blank of 150 milliseconds; then the picture was shown in the center of the screen on a white background and remained on the screen until the participant gave a verbal response through the microphone (via voice-key), either naming the picture (correctly or incorrectly) or giving an "I don't know" answer[1]. Latency and Accuracy were recorded. The Picture Naming task was built in E-Prime® and the administration lasted about 10 min.

The participants took part in the study voluntarily. The consent obtained from all participants was both written and informed.

The study was approved by the Local Ethical Committee of the School of Psychology of the University of Padua and conducted in accordance with the principles of the Declaration of Helsinki.

Statistical Analyses

A mixed-effects model approach (Pinheiro and Bates, 2000) was used. The most important advantage of this statistical approach is that it provides awareness about all factors that potentially contribute to the structure of our data (Baayen et al., 2008).

In particular, we used Generalized Linear mixed-effect models (GLMM) as a suitable method for analyzing response times and accuracy distribution (e.g., Baayen et al., 2008; Quené and Van den Bergh, 2008; Baayen and Milin, 2010). Random

[1]A limitation of our study is that the picture naming task did not register different types of errors. Name retrieval errors and recognition errors are two different responses based on different cognitive mechanisms (see Semenza, 2006 for a discussion of recognition and name retrieval errors).

FIGURE 1 | Example of figures of the Picture Naming task: famous character on the left-hand side and Logo image at the center, (https://www.istockphoto.com/ca). Object image on the right-hand side; no permission is required for the use of this image (Moreno-Martínez and Montoro, 2012).

effects for all GLMMs were ID (i.e., subject identity) and TrialList (i.e., picture identity). All Latency analyses were made on participants' correct scores. Accuracy was entered as binomial dependent variable in which the whole set of responses (4.140 data-points, both correct and incorrect) were considered in a repeated-measure design, which avoids the proportion aggregation of binomial data and provide a balanced method of analysis in psycholinguistic (Quené and Van den Bergh, 2008). Fixed effects (i.e., independent variables) for Latency and Accuracy analyses were CRI (as a continuous variable), MoCA score (as a continuous variable) adjusted for age and education, and Category (i.e., Proper Names, Logo Names and Common Nouns). All GLMM analyses started from a null model that included only an intercept; then all the independent variables were added. Likelihood Ratio Test was used for model comparison. Akaike's Information Criterion (AIC; Sakamoto et al., 1986) and Delta-AIC (Burnham and Anderson, 2003) were used to examine model plausibility. Cook's distance (Cook and Weisberg, 1982) was measured to detect influential data and next the whole dataset was considered without excluding any observations.

For the Latency analysis, CRI and MoCA were first entered separately in the null model. Next, these two variables were considered as additional terms and as interaction terms, respectively, in two separate models. The same procedure was followed for the Latency analysis considering CRI and Category as predictors of interest. See **Table 2** for more details about each model.

For the Accuracy analysis, CRI and MoCA were first entered separately as independent variables in the null model. Next, these two variables were considered as additional terms and as interaction terms, respectively, in two separate models. The same procedure was followed for Accuracy considering CRI and Category as independent variables (see **Table 3** for more details about each model).

We assessed the relationship between CRI (i.e., our predictor of interest) and psycholinguistic measures related to the pictures in order to evaluate if higher CRI was correlated with better performance in the case of items that were less familiar, more difficult and acquired later, compared to lower CRI. In addition, we assessed the relationship between Category and Psycholinguistic variables on naming performance, without entering CRI. All the analyses were performed by means of R Software (R Core Team, 2016 version 3.3.1) and GLMM was run by means of lme4-package (Bates et al., 2014), with an α level of 0.05 defining significance.

RESULTS

Results are grouped in four separate sections considering Latency and Accuracy as measures of performance in the Picture Naming task. In the first section, we report results about the effect of Cognitive Reserve on naming performance; in the second, we report the results of Cognitive Reserve and the global cognitive profile as predictors of naming performance; in the third, we report the results of Cognitive Reserve and name Category as predictors of naming performance; and finally in the fourth, we report the results of the analyses in which Cognitive Reserve, Psycholinguistic variables, and Category are entered as predictors of naming performance.

Cognitive Reserve on Naming Performance

CRI (Cognitive Reserve Index) was our measure of CR and was entered as a continuous variable. CRI significantly predicted both naming Latency ($B = -0.01$, $t = -11$; $p < 0.01$) and naming Accuracy ($B = 0.54$, $z = 3.09$; $p < 0.01$), with participants performing better when CRI was higher; see **Figure 2**.

Based on Delta-AIC as a measure to check model plausibility (Burnham and Anderson, 2003), the model with CRI as the only independent variable was 63 times more plausible than the null model for the Latency analysis, and 813 times more plausible than the null model for the Accuracy analysis.

Cognitive Reserve and Global Cognitive Profile on Naming Performance

MoCA (Montreal Cognitive Assessment, a measure of global cognitive profile) was a continuous variable entered in the null model. Based on the interaction between CRI and MoCA, these two independent variables improved the model fit and predicted name retrieval Latency ($B = 0.005$, $t = 7$; $p < 0.001$) with a stronger CRI effect when the cognitive profile score was in the low range (MoCA < 23). The model with the interaction between CRI and MoCA was about 24 times more plausible than the model with CRI as the only independent variable [χ^2

TABLE 2 | Generalized Linear Mixed-Effects Models with name retrieval Response Times as dependent variable.

	Cognitive reserve and global cognitive profile	AIC	AIC w	*Pr(>Chisq)*
Model 0	RT ~ 1 + (1|ID) + (1|TrialList)	22867.2	0.10	–
Model 1	RT ~ MoCA + (1|ID) + (1|TrialList)	22858.9	0.65	<0.001
Model 2	RT ~ CRI + (1|ID) + (1|TrialList)	22869.1	0.01	<0.001
Model 3	RT ~ CRI + MoCA + (1|ID) + (1|TrialList)	22860.9	0.24	0.024
Model 4	RT ~ CRI * MoCA + (1|ID) + (1|TrialList)	22862.7	0.09	<0.001
COGNITIVE RESERVE AND NAME CATEGORIES				
Model 1	RT ~ CAT + (1|ID) + (1|TrialList)	22864.8	0.46	<0.001
Model 3	RT ~ CRI + CAT + (1|ID) + (1|TrialList)	22866.7	0.18	<0.001
Model 4	RT ~ CRI * CAT + (1|ID) + (1|TrialList)	22867	0.15	0.15

Dependent Variable: RT (i.e., name retrieval Response Time in milliseconds. Fixed effects: MoCA, Montreal Cognitive Assessment test (Nasreddine et al., 2005); CRI, Cognitive Reserve Index (Nucci et al., 2012); CAT, name categories (i.e., Proper Names, Logo Names, and Common Nouns). Random Effects: ID, subject identity; Trials, picture identity. AIC, Akaike's Information Criterion; AIC w, AIC weight; Pr (>Chisq), Chi-Square probability associated to the model.

TABLE 3 | Generalized Linear Mixed-Effects Models with name retrieval Accuracy as dependent variable.

	Cognitive reserve and global cognitive profile	AIC	AIC w	*Pr(>Chisq)*
Model 0	ACC ~ 1 + (1|ID) + (1|TrialList)	3995.4	0.01	–
Model 1	ACC ~ MoCA + (1|ID) + (1|TrialList)	3982.1	0.01	<0.001
Model 2	ACC ~ CRI + (1|ID) + (1|TrialList)	3988.7	0.01	<0.001
Model 3	ACC ~ CRI + MoCA + (1|ID) + (1|TrialList)	3974.7	0.54	<0.001
Model 4	ACC ~ CRI * MoCA + (1|ID) + (1|TrialList)	3975.1	0.43	<0.21
COGNITIVE RESERVE AND NAME CATEGORIES				
Model 1	ACC ~ CAT + (1|ID) + (1|TrialList)	3958.1	0.01	<0.001
Model 3	ACC ~ CRI + CAT + (1|ID) + (1|TrialList)	3951.4	0.01	0.002
Model 4	ACC ~ CRI * CAT + (1|ID) + (1|TrialList)	3943	0.98	< 0.01

Dependent Variable: ACC (i.e., Accuracy of name retrieval). Fixed effects: MoCA, Montreal Cognitive Assessment test (Nasreddine et al., 2005); CRI, Cognitive Reserve Index (Nucci et al., 2012); CAT, name categories (i.e., Proper Names, Logo Names, and Common Nouns). Random Effects: ID, subject identity; Trials (picture identity). AIC, Akaike's Information Criterion; AIC w, model averaging; Pr(>Chisq), Chi-Square probability associated to the model.

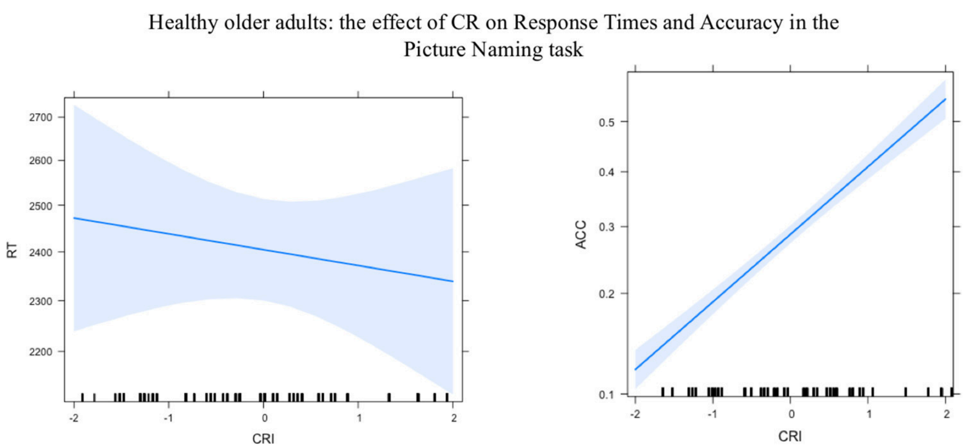


FIGURE 2 | CRI as predictor of name retrieval Response Times and Accuracy in older adults: generalized linear mixed models; Fixed effect, Cognitive Reserve Index (CRI); Dependent variables: RT, Response Times in milliseconds; ACC, Accuracy of name retrieval.

(2) = 10.35, $p < 0.01$]. This suggests that global cognitive profile and CRI taken together strongly predicted name retrieval Latency in older adults.

The MoCA score predicted the overall Accuracy of name retrieval when it was the only independent variable ($B = 0.23$, $z = 4.24$; $p < 0.01$), and adding the CRI to the MoCA improved the model fit [χ^2 (2) = 9.32, $p < 0.001$]. These two variables were then considered as interaction terms and no significant improvement was found ($B = -0.06$, $z = -1.25$; $p = 0.21$). However, the Likelihood Ratio Test carried out to compare the model with CRI and MoCA as interaction terms and the model with MoCA as a single variable showed a better fit for the former

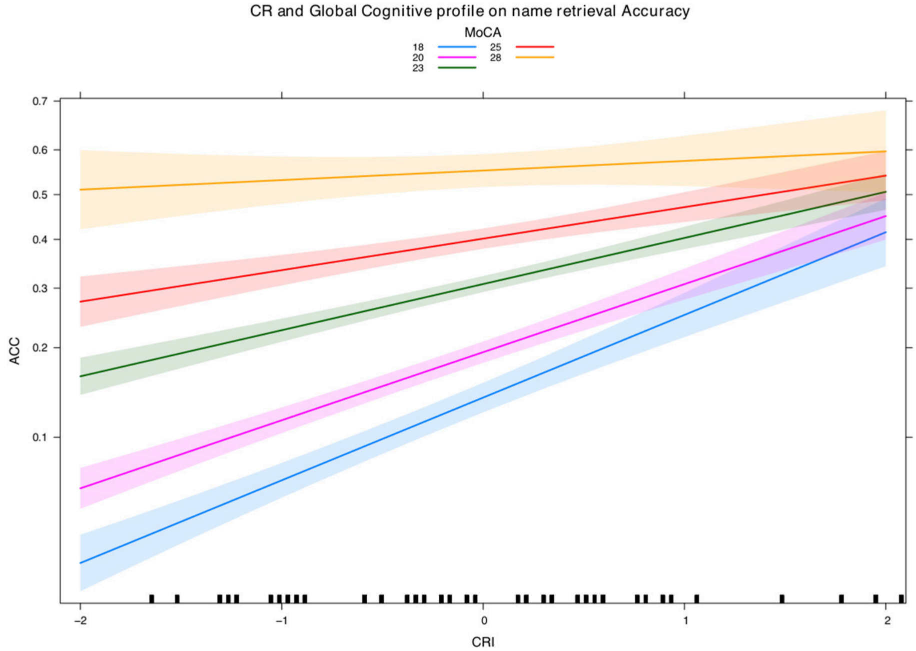


FIGURE 3 | CRI and MoCA as predictors of name retrieval Accuracy in older adults. Generalized Linear Mixed Models; MoCA, Montreal Cognitive Assessment; ACC, Accuracy; CRI, Cognitive Reserve Index.

[χ^2 (2) = 10.85, $p = 0.004$], with a stronger CRI effect when the MoCA score was in the low range (MoCA < 23; **Figure 3**).

Cognitive Reserve on Naming Performance According to Category (Proper Names, Logo Names and Common Nouns)

Entering in the null model Category as the factorial variable (three levels: Proper Names, Logo Names, and Common Nouns) significantly improved the model fit compared with the null model in both Latency [χ^2 (2) = 6.35, $p = 0.04$], and Accuracy χ^2 (2) = 41.26, $p < 0.001$). Common Nouns were retrieved faster and more accurately than Proper Names (Latency: $B = 0.15$, $t = 126$; $p < 0.001$; Accuracy: $B = -0.64$, $z = -2.05$; $p = 0.04$) and also than Logo Names (Latency: $B = 0.16$, $t = 130$; $p < 0.001$; Accuracy: $B = -2.23$, z = −6.94; $p < 0.001$). Proper Names, on the other hand, were retrieved faster and more accurately than Logo Names (Latency: $B = 0.004$, $t = 4$; $p < 0.001$; Accuracy: $B = -1.59$, $z = -4.95$; $p = 0.001$).

When Category and CRI were entered as independent variables in two separate models no effect of CRI was found on Latency across the three categories, either in the case of CRI and Category as additional terms [χ^2 (1) = 0.09, $p = 0.75$], or when CRI and Category were considered as interaction terms [χ^2 (1) = 3.75, $p = 0.15$] (see **Figure 4**).

Table 4 shows the summary result of the model of interest, (i.e., where CRI and Category are entered as interaction terms and Latency is the dependent variable).

In line with our hypothesis, Accuracy showed a significant effect of CRI across name categories. CRI and Category as interaction terms predicted Accuracy ($B = -0.35$, $z = -3.24$; $p < 0.01$) and this model was about 68 times more plausible than the model with CRI and Category entered as additional terms. Entering CRI and Category as interaction terms showed significant improvement in the model fit, compared with entering name Category as the only variable of interest [χ^2 (1) = 21.19, $p < 0.001$]. **Table 5** shows summary results of the model of interest (i.e., based on the model with CRI and Category as interaction terms and Accuracy as dependent variable).

In line with our hypothesis, the effect of CRI on name retrieval Accuracy was significantly lower for Proper Names than for Common Nouns ($B = 0.26$, $z = 2.69$, $p < 0.01$) and Logo Names ($B = 0.35$, $z =$ 3.24, $p < 0.01$). The effect of CRI on Accuracy did not differ when comparing Logo Names and Common Nouns ($B = -0.09$, $z = -0.81$, $p = 0.41$) (**Figure 5**).

Cognitive Reserve and Psycholinguistic Variables

Psycholinguistic variables showed a different role in the three categories on the Accuracy. Participants' performance in the picture naming task was better for higher familiar pictures ($B = 0.38$, $z = 3.05$, $p < 0.01$) and for those whose names that were easier to retrieve ($B = 0.36$, $z = -2.37$, $p = 0.01$), with no significant effect of Age of Acquisition ($B = 0.27$, $z = 1.82$, $p = 0.06$). *Post-hoc* analyses showed that Familiarity predicted Accuracy in Proper Names ($B = 0.66$, $z = 2.42$, $p = 0.01$), but not in Common Nouns ($B = -0.03$, $z = -0.24$, $p = 0.81$) and not in Logo Names ($B = 0.03$, $z = 0.24$, $p = 0.81$). In a similar way, Difficulty predicted Accuracy in Proper Names ($B = -4.58$, $z = -2.91$, $p < 0.01$), but not in Common Nouns

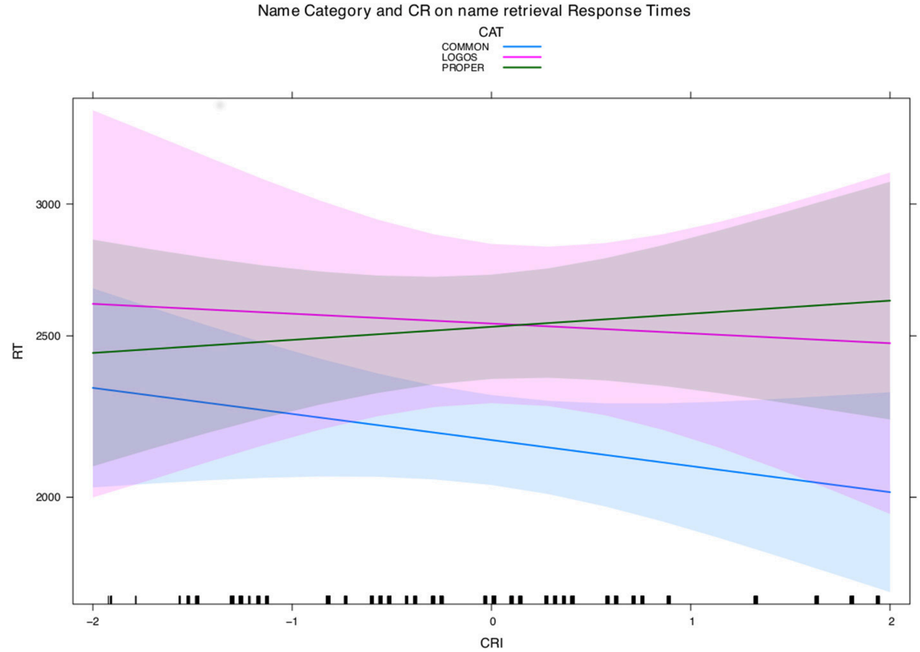


FIGURE 4 | CRI and Category as predictors of name retrieval Latency in older adults: Generalized Linear Mixed Models results. CRI, Cognitive Reserve Index questionnaire; CT, category; RT, name retrieval speediness in milliseconds.

TABLE 4 | Generalized Linear Mixed-Effects Models results for picture naming Response Times, in a group of healthy older adults.

	CR and category (Latency)			
	Beta	**SE**	**Z**	***p*-value**
(Intercept)	7.68	0.07	105.58	<0.001
LN	0.16	0.07	2.14	0.03
PN	0.15	0.07	2.23	0.02
CRI	−0.03	0.04	−0.75	0.45
CRI: LN-CN	0.02	0.03	−0.56	0.57
CRI:PN-CN	−0.05	0.02	−1.94	0.05
CRI:PN-LN	−0.03	0.04	0.76	0.44

Dependent variable: Latency for name retrieval (milliseconds). Fixed effects: CRI, Cognitive Reserve Index and name Category (CN, Common Nouns; PN, Proper Names; LN, Logo Names). Random effects: subject identity and trial list. CRI:PN-CN was obtained in the relevel function in R.

TABLE 5 | Generalized Linear Mixed-Effects Models results for picture naming Accuracy, in a group of healthy older adults.

	CRI and category (Accuracy)			
	Beta	**SE**	**Z**	***p*-value**
(Intercept)	−0.58	0.28	−2.06	0.03
CN	0.62	0.31	1.97	< 0.4
LN	−1.66	0.32	−5.11	< 0.001
CRI	0.36	0.18	1.97	0.4
CRI:LN-PN	0.35	0.11	3.24	< 0.01
CRI:NC-PN	0.26	0.09	2.69	< 0.01
CRI:NC-LN	−0.09	0.11	−0.81	0.41

Dependent variable: Accuracy of name retrieval. Fixed effects: CRI, Cognitive Reserve Index and name Category (CN, Common Nouns; PN, Proper Names; LN, Logo Names). CRI:NC-LN was obtained by the relevel function in R. Random effects: subject identity and trial list.

($B = 0.98$, $z = 1.04$, $p = 0.29$) and not in Logo Names ($B = 0.96$, $z = 1.03$, $p = 0.31$).

We evaluated if higher CR predicted Accuracy of Names that were less familiar, more difficult to retrieve and acquired later. Entering CRI and Familiarity as interaction terms predicted Accuracy ($B = -0.002$, $z = -1.96$, $p = 0.04$), showing that when Familiarity was low, the name retrieval performance of participants with higher CRI was better. Similarly, CRI and Difficulty as interaction terms significantly predicted Accuracy ($B = -0.003$, $z = -2.29$, $p = 0.02$), whereas no interaction was found between CRI and Age of acquisition ($B = 0.003$, $z = 0.81$, $p = 0.41$).

DISCUSSION

This research aimed to investigate whether language, in particular name retrieval processing, can be influenced by the degree of CR acquired throughout the lifespan. Some evidence of beneficial effects of CR in the domain of language has already been reported in previous studies. For instance, years of education has been shown to predict verbal comprehension skills (Schaie, 1989) and the amount of vocabulary in adulthood (Christensen et al., 1997; Arbuckle et al., 1998). However, more recent studies have reported that the effect of CR is not evident in some other cases, as in proper name retrieval tasks (Montemurro et al., 2018).

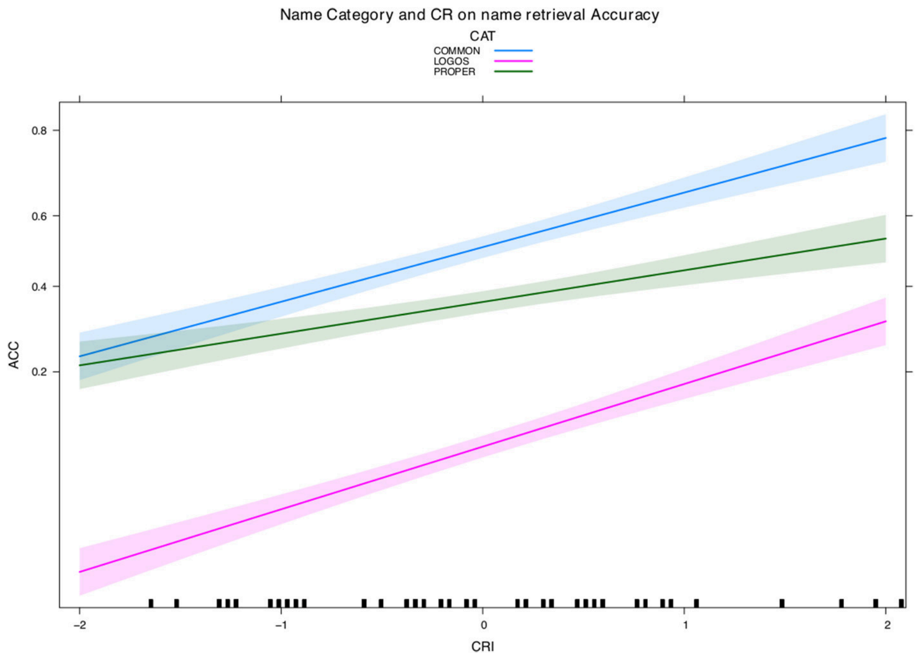


FIGURE 5 | CRI and category as predictors of name retrieval Accuracy in older adults: Generalized Linear Mixed-Model results. CRI, Cognitive Reserve Index; CAT, name category; ACC, Accuracy.

Among the group of older adults that took part in this study, those with higher CR performed better in name retrieval compared with people with lower CR. A recent study (Montemurro et al., 2018) reported that higher CR is associated with the more successful global cognitive performance of older adults with early symptoms of dementia, but not of healthy controls. Consistent with that, the present study shows that CR can affect naming performance depending on the degree of cognitive decline. In particular, name retrieval of participants with early signs of decline seem to have benefitted more from CR, than healthy participants. This result could be explained by the two mechanisms of neural reserve and neural compensation (Barulli and Stern, 2013), where the strong association between CR and name retrieval in case of cognitive decline may reflect the necessity to recruit additional networks to cope with a relatively simple task.

Evidence from previous research based on a comparison between younger and older adults has shown that as the task load increases, healthy older adults recruit brain networks in the same way as their younger counterpart (Ansado et al., 2013). The authors suggested that older adults may first adopt cognitive compensatory mechanisms and then, when compensatory processes are not enough to cope with an increased task demand, healthy older adults make use of their neural reserve (Ansado et al., 2013). Interestingly, their findings suggested that, in older adults, the neural substrates of CR are based on flexible and adaptive neural processes; however, such adaptation of neural responses can be more or less successful depending both on global cognitive condition and the difficulty of the task.

In the present study we used a simple name retrieval task, which might require additional resources only in persons with an impaired global condition, whereas such additional resources might be needed in healthy older adults only if task demand got higher. We analyzed the role of CR in name retrieval, which is frequently impaired in adults who often refer to this problem in terms of "memory loss." Difficulties in retrieving names can be one of the first symptoms reported at the early clinical assessments and may generally depend on a physiological decline (e.g., MacKay and Burke, 1990; Rastle and Burke, 1996; Semenza, 2006). In our study we reported that naming performance across categories showed Logo Names as the most demanding items, but when entering CR in the model, high CR predicted a better name retrieval performance both in the Logo Name and the Common Noun categories. In naming Proper Names, high vs. low CR did not influence performance as it did in the two other name categories. In other words, CR barely modulates Proper name retrieval; it is correlated with better performance in retrieving Common Nouns and Logo Names, which may be both conceptually categorized due to their greater environmental pervasiveness, rather than naming with Proper Names, which are pure referential expressions. Thus, these results underline that CR matters for naming performance only in some cases.

To the best of our knowledge, this is the first study exploring the relationship between CR and name retrieval using the three name categories of proper Names, Logo Names, and Common Nouns. The interest in this relationship derives from previous findings, where proper name anomia was proposed to be a predictor for the onset of dementia (Semenza et al., 2003). In their study, the authors suggested that proper name anomia at the very early stage of dementia might be due to lexical semantic disruption. Although the impact of age-related naming deficits has already been reported in previous studies (e.g., Flicker et al., 1987; Evrard, 2002), the contribution of CR on name retrieval processing in older adults has been addressed only recently (e.g., Mondini and Semenza, 2016; Montemurro et al., 2018).

Proper Names and Common Nouns were considered in light of the well-documented dissociation between proper name and common noun retrieval in older adults (e.g., Cohen and Burke, 1993; Rastle and Burke, 1996; Evrard, 2002; Tsukiura et al., 2011), even if the evidence of a disproportionate deficit for proper names as compared to common nouns has been shown to be controversial (see Maylor, 1997). In line with our hypothesis, we demonstrated that the effect of CR on name retrieval Accuracy was weaker for Proper Names compared to Logo Names and Common Nouns. This result seems to be reflected in the analysis of Response Times (see **Figure 4**), although no statistical difference across name categories was found when considering CR as a predictor.

Common Nouns and Logo Names could highly benefit from CR, as shown in previous studies where CR was associated with semantic task components (see Reed et al., 2011; Darby et al., 2017). Our results showed differences across the three categories for Accuracy but not for Latency, possibly because name retrieval speediness is more generally affected by age-related synaptic delay (MacKay and Burke, 1990; Rastle and Burke, 1996; Jackson et al., 2012). The difference between Accuracy and Latency might be due to high age-related variability when performing a cognitive task that requires reaction speediness (e.g., Anstey, 1999; Christensen et al., 1999; Hultsch et al., 2002; Bielak et al., 2010). For example, in a longitudinal study employing a series of cognitive tasks conducted with 760 elderly individuals, Christensen et al. (1999) found a heterogeneous pattern of speed performance along with increased age-related inter-individual variation.

Our results show that high CR is a predictor of name retrieval Accuracy in the case of less familiar and more difficult names. However, no association was found between CRI and Age of Acquisition. The same result was obtained when excluding from the analysis Logo Names, which were generally rated as acquired later in life. This suggests that name retrieval accuracy does not appear to be sensitive to CR at the point in time when names were acquired, and that CR could be associated with the frequency of occurrence of certain names throughout lifespan experiences. Interestingly, our analysis of the association between psycholinguistic variables and Accuracy across categories shows that Proper Names can be more sensitive to image Familiarity and Difficulty of naming, compared to Common Nouns and Logo Names. These findings suggest that name retrieval Accuracy seems to be modulated not only by Familiarity and Difficulty, but also by CR. Categories, however, showed that name retrieval of those Names that refer to repeated and shared knowledge (i.e., Common Nouns and Logo Names) is better with higher CR, whereas name retrieval of items with poor semantic attributes (i.e., Proper Names) benefit from high image Familiarity and low Difficulty of naming.

Thus, it might be the case that higher CR can help context-driven information processing, as in the case of real living/non-living entities and commercial products, instead of information that are arbitrarily assigned to unique entities, despite their familiarity, or ease of accessing their representation. With this interpretation, context availability might be a very interesting target to explore in the future, in association with the possible effects of CR.

In sum, our results showed that CR predicts overall name retrieval performance and that this effect seems to be higher when the global cognitive profile shows very early symptoms of age-related cognitive decline. Additionally, CR may act as a stronger predictor of name retrieval when names are involved in denser semantic networks (i.e., Logo Names and Common Nouns) compared to lexical labels that are pure referential expressions (i.e., Proper Names). These results suggest that CR, which relies on socially active life-style and exposure to world knowledge, can modulate cognitive performance. On the other hand, our observation of a weak dependence between CR and some cognitive tasks, such as Proper Name retrieval, may make the latter especially useful to detect early symptoms of dementia in individuals with high CR.

Future investigations might further address the proposition that having an active and socially integrated life-style in adulthood may not only enrich cognitive resources in general, but also strengthen some specific cognitive processes, such as name retrieval.

ETHICS STATEMENT

COMITATO ETICO DELLA RICERCA PSICOLOGICA (AREA 17) Dipartimenti/Sezione di Psicologia—Universitàdi Padova, Via Venezia 8, 35131, Padova, FAX. +39-0498276600, e-mail: comitato.etico17@gmail.com; Sito WEB: http://ethos.psy.unipd.it/

Protocollo: 2372 (SOSTITUISCE IL N.2324)

Data: 02/08/2017

Numero Univoco: 8696E7DA11D7B45656D6DDCE6CD673B0

Scopo: Richiesta di parere

Titolo: Riserva cognitiva e velocità di denominazione nell'anziano sano e con decadimento cognitivo lieve-moderato: uso di un test computerizzato.

nomi propri nel declino cognitivo.

Proponente

Cognome e nome: Mondini Sara Ruolo: Associato

e-mail: sara.mondini@unipd.it

Area: Psicologia generale (se altro):

Ricercatori partecipanti: 3

SaM (Docente presso il Dipartimento di Psicologia Generale), Sonia Montemurro (Dottoranda presso il Dipartimento di Psicologia Generale), Chiara Crovace (Studentessa presso il Dipartimento di Psicologia Generale).

Il Comitato Etico, dopo attento esame delle informazioni fornite dal proponente, esprime parere positivo riguardante gli aspetti etici del progetto.

The project has been approved by the Ethical Committee for the Psychological Research of the University of Padova.

Padova, 11/10/2017

AUTHOR CONTRIBUTIONS

SoM: devised the experiment, analyzed the data, and wrote the manuscript. SaM: devised the experiment, discussed data and wrote the manuscript. CC: collected the data. GJ: discussed and wrote the manuscript.

ACKNOWLEDGMENTS

We thank Luca Semenzato for his technical assistance.

REFERENCES

Almond, N. M., and Morrison, C. (2017). Effects of aging and recall of common and uncommon first names using the face-name association technique compared with the pure-lists technique over repeated trials. *Health. Aging Res.* 6:e4. doi: 10.1097/HXR.0000000000000004

Ansado, J., Monchi, O., Ennabil, N., Deslauriers, J., Jubault, T., Faure, S., et al. (2013). Coping with task demand in aging using neural compensation and neural reserve triggers primarily intra-hemispheric-based neurofunctional reorganization. *Neurosci. Res.* 75, 295–304. doi: 10.1016/j.neures.2013.01.012

Anstey, K. J. (1999). Sensorimotor variables and forced expiratory volume as correlates of speed, accuracy, and variability in reaction time performance in late adulthood. *Aging Neuropsychol. Cogn.* 6, 84–95. doi: 10.1076/anec.6.2.84.786

Arbuckle, T. Y., Maag, U., Pushkar, D., and Chaikelson, J. S. (1998). Individual differences in trajectory of intellectual development over 45 years of adulthood. *Psychol. Aging* 13:663. doi: 10.1037/0882-7974.13.4.663

Arcara, G., Mondini, S., Bisso, A., Palmer, K., Meneghello, F., and Semenza, C. (2017). The relationship between cognitive reserve and math abilities. *Front. Aging Neurosci.* 9:429. doi: 10.3389/fnagi.2017.00429

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Baayen, R. H., and Milin, P. (2010). Analyzing reaction times. *Int. J. Psychol. Res.* 3, 12–28. doi: 10.21500/20112084.807

Barulli, D., and Stern, Y. (2013). Efficiency, capacity, compensation, maintenance, plasticity: emerging concepts in cognitive reserve. *Trends Cogn. Sci.* 17, 502–509. doi: 10.1016/j.tics.2013.08.012

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv:* 1406.5823. doi: 10.18637/jss.v067.i01

Bielak, A. A., Hultsch, D. F., Strauss, E., MacDonald, S. W., and Hunter, M. A. (2010). Intraindividual variability is related to cognitive change in older adults: evidence for within-person coupling. *Psychol. Aging* 25:575. doi: 10.1037/a0019503

Brédart, S., Brennen, T., and Valentine, T. (1997). Dissociations between the processing of proper and common names. *Cogn. Neuropsychol.* 14, 209–217. doi: 10.1080/026432997381556

Burke, D. M., MacKay, D. G., Worthley, J. S., and Wade, E. (1991). On the tip of the tongue: what causes word finding failures in young and older adults? *J. Memory Lang.* 30, 542–579. doi: 10.1016/0749-596X(91)90026-G

Burnham, K. P., and Anderson, D. R. (2003). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York, NY: Springer Science and Business Media.

Christensen, H., Korten, A. E., Jorm, A. F., Henderson, A. S., Jacomb, P. A., Rodgers, B., et al.. (1997). Education and decline in cognitive performance: compensatory but not protective. *Int. J. Geriatr. Psychiatry* 12, 323–330.

Christensen, H., Mackinnon, A. J., Korten, A. E., Jorm, A. F., Henderson, A. S., and Jacomb, P. (1999). Dispersion in cognitive ability as a function of age: a longitudinal study of an elderly community sample. *Aging Neuropsychol. Cogn.* 6, 214–228. doi: 10.1076/anec.6.3.214.779

Cohen, G., and Burke, D. M. (1993). Memory for proper names: a review. *Memory* 1, 249–263. doi: 10.1080/09658219308258237

Conti, S., Bonazzi, S., Laiacona, M., Masina, M., and Coralli, M. V. (2015). Montreal Cognitive Assessment (MoCA)-Italian version: regression based norms and equivalent scores. *Neurol. Sci.* 36, 209–214. doi: 10.1007/s10072-014-1921-3

Cook, T. D., and Weisberg, S. (1982). *Residuals and Influence in Regression*. New York, NY: Chapman and Hall.

Darby, R. R., Brickhouse, M., Wolk, D. A., and Dickerson, B. C. (2017). Effects of cognitive reserve depend on executive and semantic demands of the task. *J. Neurol. Neurosurg. Psychiatr.* 88, 794–802. doi: 10.1136/jnnp-2017-315719

Erber, J. T., and Szuchman, L. T. (1996). Memory performance in relation to age, verbal ability, and activity. *Exp. Aging Res.* 22, 59–72. doi: 10.1080/03610739608253997

Evrard, M. (2002). Ageing and lexical access to common and proper names in picture naming. *Brain Lang.* 81, 174–179. doi: 10.1006/brln.2001.2515

Feyereisen, P. (1997). A meta-analytic procedure shows an age-related decline in picture naming: comments on Goulet, Ska, and Kahn (1994). *J. Speech Lang. Hear. Res.* 40, 1328–1333. doi: 10.1044/jslhr.4006.1328

Flicker, C., Ferris, S. H., Crook, T., and Bartus, R. T. (1987). Implications of memory and language dysfunction in the naming deficit of senile dementia. *Brain Lang.* 31, 187–200. doi: 10.1016/0093-934X(87)90069-1

Gontijo, P. F., Rayman, J., Zhang, S., and Zaidel, E. (2002). How brand names are special: brands, words, and hemispheres. *Brain Lang.* 82, 327–343. doi: 10.1016/S0093-934X(02)00036-6

Gorno-Tempini, M. L.,Price, C. J., Josephs, O., Vandenberghe, R., Cappa, S. F., Kapur, N., et al.. (1998). The neural systems sustaining face and proper-name processing. *Brain* 121, 2103–2118. doi: 10.1093/brain/121.11.2103

Greene, J. D., and Hodges, J. R. (1996). The fractionation of remote memory: evidence from a longitudinal study of dementia of Alzheimer type. *Brain* 119, 129–142. doi: 10.1093/brain/119.1.129

Hultsch, D. F., Hammer, M., and Small, B. J. (1993). Age differences in cognitive performance in later life: Relationships to self-reported health and activity life style. *J. Gerontol.* 48, P1–P11. doi: 10.1093/geronj/48.1.P1

Hultsch, D. F., MacDonald, S. W., and Dixon, R. A. (2002). Variability in reaction time performance of younger and older adults. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 57, P101–P115. doi: 10.1093/geronb/57.2.P101

Jackson, J. D., Balota, D. A., Duchek, J. M., and Head, D. (2012). White matter integrity and reaction time intraindividual variability in healthy aging and early-stage Alzheimer disease. *Neuropsychologia* 50, 357–366. doi: 10.1016/j.neuropsychologia.2011.11.024

Kripke, S. (1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

MacKay, D. G., and Burke, D. M. (1990). "Cognition and aging: a theory of new learning and the use of old connections," in *Aging and Cognition: Knowledge Organization and Utilization Psychology*, ed T. M. Hess (North-Holland: North-Holland Press), 213–263. doi: 10.1016/S0166-4115(08)60159-4

Maruish, M. E., and Moses, J. A. (1996). *Clinical Neuropsychology: Theoretical Foundations for Practitioners*. New York, NY: Psychology Press.

Maylor, E. A. (1997). Proper name retrieval in old age: converging evidence against disproportionate impairment. *Aging Neuropsychol. Cogn.* 4, 211–226. doi: 10.1080/13825589708256648

Mondini, S., and Semenza, C. (2016). "Cognitive reserve and ageing. What does cognitive reserve protect in ageing? Frontiers in Psychology," in *Abstract Retrieved From 54th Annual Academy of Aphasia Meeting*. Llandudno, UK.

Montemurro, S., Mondini, S., Nucci, M., and Semenza, C. (2018). Proper name retrieval in cognitive decline. *Ment. Lex.* 13, 215–229. doi: 10.1075/ml.18004.mon

Moreno-Martínez, F. J., and Montoro, P. R. (2012). An ecological alternative to Snodgrass and Vanderwart: 360 high quality colour images with norms for seven psycholinguistic variables. *PLoS ONE 7*:e37527. doi: 10.1371/journal.pone.0037527

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., et al. (2005). The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* 53, 695–699. doi: 10.1111/j.1532-5415.2005.53221.x

Nucci, M., Mapelli, D., and Mondini, S. (2012). Cognitive Reserve Index questionnaire (CRIq): a new instrument for measuring cognitive reserve. *Aging Clin. Exp. Res.* 24, 218–226. doi: 10.3275/7800

Pelamatti, G., Pascotto, M., and Semenza, C. (2003). Verbal free recall in high altitude: proper names vs common names. *Cortex* 39, 97–103. doi: 10.1016/S0010-9452(08)70077-7

Pinheiro, J. C., and Bates, D. M. (2000). "Linear mixed-effects models: basic concepts and examples," in *Mixed-Effects Models in S and S-PLUS. Statistics and Computing*. (New York, NY: Springer).

Quené, H., and Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *J. Mem. Lang.* 59, 413–425. doi: 10.1016/j.jml.2008.02.002

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/

Rastle, K. G., and Burke, D. M. (1996). Priming the tip of the tongue: effects of prior processing on word retrieval in young and older adults. *J. Mem. Lang.* 35, 586–605. doi: 10.1006/jmla.1996.0031

Reed, B. R., Dowling, M., Farias, S. T., Sonnen, J., Strauss, M., Schneider, J. A., et al. (2011). Cognitive activities during adulthood are more important than education in building reserve. *J. Int. Neuropsychol. Soc.* 17, 615–624. doi: 10.1017/S1355617711000014

Robertson, I. H. (2013). A noradrenergic theory of cognitive reserve: implications for Alzheimer's disease. *Neurobiol. Aging* 34, 298–308. doi: 10.1016/j.neurobiolaging.2012.05.019

Roldán-Tapia, L., García, J., Cánovas, R., and León, I. (2012). Cognitive reserve, age, and their relation to attentional and executive functions. *Appl. Neuropsychol. Adult* 19, 2–8. doi: 10.1080/09084282.2011.595458

Sakamoto, Y., Ishiguro, M., Kitagawa, G. (1986). *Akaike Information Criterion Statistics*. Dordrecht: D. Reidel.

Salmon, J. P., McMullen, P. A., and Filliter, J. H. (2010). Norms for two types of manipulability (graspability and functional usage), familiarity, and age of acquisition for 320 photographs of objects. *Behav. Res. Methods* 42, 82–95. doi: 10.3758/BRM.42.1.82

Scarmeas, N., Levy, G., Tang, M. X., Manly, J., and Stern, Y. (2001). Influence of leisure activity on the incidence of Alzheimer's disease. *Neurology* 57, 2236–2242. doi: 10.1212/WNL.57.12.2236

Schaie, K. W. (1989). The hazards of cognitive aging. *Gerontologist* 29, 484–493. doi: 10.1093/geront/29.4.484

Searle, J. (1969). *Speech Acts*. Cambridge: Cambridge University Press.

Semenza, C. (2006). Retrieval pathways for common and proper names. *Cortex* 42, 884–891. doi: 10.1016/S0010-9452(08)70432-5

Semenza, C. (2009). The neuropsychology of proper names. *Mind Lang.* 24, 347–369. doi: 10.1111/j.1468-0017.2009.01366.x

Semenza, C., Mondini, S., Borgo, F., Pasini, M., and Sgaramella, M. T. (2003). Proper names in patients with early Alzheimer's disease. *Neurocase* 9, 63–69. doi: 10.1076/neur.9.1.63.14370

Semenza, C., and Zettin, M. (1989). Evidence from aphasia for the role of proper names as pure referring expressions. *Nature* 342:678. doi: 10.1038/342678a0

Stern, Y. (2006). Cognitive reserve and Alzheimer disease. *Alzheimer Dis. Assoc. Disord.* 20, S69–S74. doi: 10.1097/01.wad.0000213815.20177.19

Stern, Y. (2012). Cognitive reserve in ageing and Alzheimer's disease. *Lancet Neurol.* 11, 1006–1012. doi: 10.1016/S1474-4422(12)70191-6

Tsukiura, T., Sekiguchi, A., Yomogida, Y., Nakagawa, S., Shigemune, Y., Kambara, T., et al. (2011). Effects of aging on hippocampal and anterior temporal activations during successful retrieval of memory for face–name associations. *J. Cogn. Neurosci.* 23, 200–213. doi: 10.1162/jocn.2010.21476

Valentine, V. M. T. (1998). The effect of age of acquisition on speed and accuracy of naming famous faces. *Q. J. Exp. Psychol. A* 51, 485–513. doi: 10.1080/713755779

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Communication

ORIGINAL RESEARCH
published: 14 May 2019
doi: 10.3389/fcomm.2019.00016

Check for updates

Aging and Language: Maintenance of Morphological Representations in Older Adults

Phaedra Royle[1,2]*, *Karsten Steinhauer*[2,3], *Émie Dessureault*[1], *Alexandre C. Herbay*[2,3] and *Simona M. Brambati*[4]

[1] *École d'orthophonie et d'audiologie, Université de Montréal, Montreal, QC, Canada,* [2] *Centre for Research on Brain Language and Music, Montreal, QC, Canada,* [3] *School of Communication Sciences and Disorders, McGill University, Montreal, QC, Canada,* [4] *Département de Psychologie, Université de Montréal, Montreal, QC, Canada*

Studies employing primed lexical decision tasks have revealed morphological facilitation effects in children and young adults. It is unknown if this effect is preserved or diminished in older adults. In fact, only few studies have investigated age-related changes in morphological processing and results are inconsistent across studies. To address this issue, we investigated inflection morphology compared to orthographic and semantic processing in young and older adults. Twenty-six adults aged 60–85 and 22 younger adults aged 19–28 participated. We probed verb recognition using a sandwich-masked primed lexical decision paradigm. We investigated lexical decision using different prime presentation times (PPTs) (33, 66, and 150 ms), and prime types with priming conditions involving orthographic (e.g., *cassis—CASSE* 'blackcurrant—break'), regular inflection morphological (*cassait—CASSE* 'broke—break'), and semantic primes (*brise—CASSE* 'break—break') and their controls, while measuring response accuracy and reaction times. Response accuracy analyses revealed that older participants performed at ceiling on the lexical decision task, and that accuracy levels were higher compared to young adults. Reaction-time data revealed effects of age group, priming condition, and an interaction of age group and morphological priming, but no PPT effects. Both young and older adults presented a significant facilitation effect (reduced reaction times) in the orthographic and morphological priming conditions. No semantic effects were observed in either group. Younger adults also showed a significantly stronger morphological priming effect, while older adults showed no difference between orthographic and morphological priming when comparing priming magnitudes. These findings suggest (1) that regular inflectional morphological processing benefits lexical access in younger French adults, confirming studies in other languages, and (2) that this advantage is reduced at older ages.

Keywords: aging, inflection morphology, masked priming, lexical decision, lexical semantics, orthographic processing, French

OPEN ACCESS

Edited by:
Juhani Järvikivi,
University of Alberta, Canada

Reviewed by:
Jana Reifegerste,
Georgetown University, United States
Raymond Bertram,
University of Turku, Finland

Correspondence:
Phaedra Royle
phaedra.royle@umontreal.ca

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *18 April 2018*
Accepted: *02 April 2019*
Published: *14 May 2019*

Citation:
Royle P, Steinhauer K, Dessureault É, Herbay AC and Brambati SM (2019) Aging and Language: Maintenance of Morphological Representations in Older Adults. Front. Commun. 4:16. doi: 10.3389/fcomm.2019.00016

INTRODUCTION

This article addresses two objectives. Firstly, it aims to investigate the role of inflectional morphological representation in the French mental lexicon. Secondly, it aims to determine the effect of age on these morphological representations. Age has an impact on some aspects of language processing, in particular lexical access. This is anecdotally observed in everyday life. Older adults often manifest and report word-finding difficulties in spontaneous speech. These difficulties are confirmed by studies investigating characteristics of older adults' speech (Shewan and Henderson, 1988; Kemper et al., 2001; Kavé and Goral, 2017). This profile raises the question of whether lexical access difficulties are caused by word-representation degradation in the mental lexicon. While it is generally assumed that semantic aspects of word representations are preserved in older adults, little is known of age effects on morphology. This is a crucial issue for a better understanding of the nature of language difficulties encountered by normally-aging older adults in everyday life. To address this question, we deployed a masked primed lexical-decision (LD) paradigm to investigate the relative impact of orthography, semantics, and morphology on word recognition in healthy French-speaking young and older adults.

In linguistics, morphology refers to the internal structure of complex words (e.g., *in-divis-ible*). Most words that we encounter are morphologically complex, i.e., they consist of at least two morphemes (e.g., *indivisible*, contains *in-*, *divis-* and *-ible*). Morphemes can be described as minimal language units that possess a relatively consistent phonological or orthographic form and carry a part of the word's semantic or syntactic information. Many psycholinguists assume that morphemes and their meaning, rather than complex words, are stored in the mental lexicon, our long-term memory for words, and processed online. Thus, in order to understand a morphologically complex word, a listener's mind/brain has to decompose it into its morphemes and look up their corresponding information in the mental lexicon. Furthermore, morphemes can be recombined with other morphemes to create new words (Aronoff and Fudemann, 2011). From this perspective, morphological structure is an independent level of linguistic representation that needs to be processed online during language production and comprehension (see Amenta and Crepaldi, 2012, for a review). Some propose that morphological processing is obligatory and irrepressible (Stockall and Marantz, 2006). However, other researchers consider that words are processed as chunks, or whole words (e.g., *indivisible*), and that linking between the orthographic or phonological pattern of a word and its meaning does not require morphological representations or processing (Bates and Goodman, 1997; Seidenberg and Gonnerman, 2000; Devlin et al., 2004). According to this eliminativist stance, morphology is epiphenomenal to phonology/orthography or semantics and has no independent role to play in lexical representation and processing. Other more recent models assume that it is not necessary to postulate morphological, phonological or even semantic relations between words (see e.g., Baayen et al., 2011), and modeling morphological effects is processed via a symbolic layer of orthographic nodes and a symbolic layer of meanings, while using cue-based learning to predict outcomes and to learn (Milin et al., 2017). Note that there are hybrid (often termed *dual-route*) models where lexical processing can be both whole-word or morphological (e.g., high-frequency words may develop whole-word representations; Alegre and Gordon, 1999), but these frameworks are largely compatible with morphological models, as they also allow for morphemic representation (see e.g., Frauenfelder and Schreuder, 1992). In sum, the status of morphology is controversial across current theories of word recognition and production.

Why is morphology important? There is accumulating evidence that morphology impacts language acquisition and processing. Morphological awareness is important in learning language and writing development (Bertram et al., 2001; Pacton and Deacon, 2008; Quémart et al., 2011; Rvachew et al., 2017; Marquis and Royle, 2019). Given its role in language acquisition, we can wonder whether this ability is maintained during aging. Aging adults show heterogeneous patterns of cognitive abilities with some decline but also strengths relative to younger adults (see below and Ansaldo et al., 2013). Because morphology is implicated in lexical access, which can deteriorate with age, it is conceivable that morphological processing may also deteriorate with age. However, with respect to aging, the data on the impact of morphology in word reading remains rare and inconclusive. We review some studies focusing on morphological processing and aging below, but note that they are few and far between. First we discuss a classic research paradigm deployed for the study of morphology: primed lexical decision. This approach has the added value of allowing us to distinguish between different accounts of morphological processing, i.e., classical morphological representation vs. eliminativist approaches.

Priming Studies

Priming paradigms in conjunction with word processing tasks, such as lexical decision or go-no-go naming, are powerful tools to study the organization of the mental lexicon. Priming refers to a facilitation effect (or inhibition, when the effect is negative) in target stimulus processing, induced by the prior presentation of a related item (the prime). Participants' reaction times (RTs) and response accuracies (RAs) can be modulated by prime-target relationships. Target recognition facilitation in the form of decreased reaction times (RTs) have been reported for semantically related prime-target word pairs (such as *doctor-nurse,* Neely et al., 1989). Facilitation effects have also been reported when prime-target word pairs are orthographically or phonologically related (e.g., *HELP-helm*, Bijeljac-Babic et al., 1997). These findings have been interpreted as suggesting that the mental lexicon's organization is based on orthographic or phonological as well as semantic properties of words.

Priming studies have investigated whether morphological structure is an essential level of linguistic representation within the mental lexicon. Morphological priming has been investigated using different prime types, such as pseudo-derivational (e.g., *corner-corn,* where the pseudo-stem *corn*, is not part of *corner*'s morphological structure), derivational (e.g., *trucker-truck*), or inflectional priming (e.g., *vowed-vow*). The majority of studies have focused on derivational priming (see e.g., Rastle and Davis, 2008; see Amenta and Crepaldi, 2012, for a review). However, we were interested in testing whether inflectional morphological processes have potential to facilitate lexical processing. Raveh and Rueckl (2000) shows that inflected and derived primes induce similar priming effects on stem recognition at 50 ms stimulus-onset asynchrony (SOA, the total amount of time between prime and target presentation, in this case a prime presentation time (PPT) of 40 ms plus 10 ms backward mask), while inflected words induced more priming than derivation at longer SOAs (150 and 250 ms). Furthermore, inflectional morphological priming has the advantage of encoding regular semantic relationships, allowing for the study of highly constrained morphological effects (Royle et al., 2012), contrary to derivational morphological processes, which have variable semantic transparency, that is they can be opaque (e.g., *arch–archer*), but can also change lexical category (e.g., *to sail–sailor*). A study by Feldman and Prostko (2002) compared different types of priming on verb recognition in English. More specifically, they compared unmasked orthographically (and phonologically) related priming (e.g., *vowel—vow*) to semantic (e.g., *promise—vow*) and inflectional morphological priming (e.g., *vowed—vow*), in a number of different tasks and with different PPTs, ranging from barely perceptible primes (33 ms) to reliably perceptible (116 ms) and longer (300 ms). This type of manipulation allows one to investigate the time-course of different priming mechanisms and their respective decay during word processing. Overall, their data indicate that morphological priming is always facilitating (resulting in shorter RTs), while semantic and orthographic priming can be facilitating or not (or even inhibitory) depending on PPT and task. For instance, semantically oriented tasks, such as lexical decision and go-no-go naming, promote semantic priming effects, while orthographic effects are reduced or even inhibitory when the prime becomes perceptible. However, Feldman and Prostko interpret their data as being "[...] more compatible with a dynamic account where morphological effects *emerge* from conjoint influences of orthographic and semantic similarity that stabilize over time." (p. 25, our italics), that is, they promote an eliminitavist approach where so-called morphological effects are in fact the result of combined orthographic and semantic facilitation. If this is the case, morphological priming effects are expected to vary as a function of the orthographic (or phonological) and semantic overlap between the prime and target. Recall however, that Feldman and Prostko (2002) find orthographic facilitation during their priming task only at 66 ms, and semantic priming only at long PPTs (116 and 300 ms), while morphological facilitation was significant across all PPTs, suggesting that morphological priming cannot be explained by a combination of these two effects, which in turn makes an eliminativist interpretation for morphological priming hard to maintain. An important aspect of PPT manipulations is that they are assumed to constrain or highlight different cognitive processes underlying lexical access. Effects arising at shorter PPTs or processing times (SOAs) are assumed to reflect automatic and less strategic processing, while those emerging after longer PPTs are assumed to reflect less automatic, and more strategic, or even post-lexical processing (see e.g., McKoon and Ratcliff, 1995; Steinhauer et al., 2017). Primes are typically perceived at about 60 ms PPT, without participants being able to identify what they are (Forster, 1998). However some participants can be more or less attuned to primes at these PPTs, depending on their inherent processing abilities or whether masking is used (Brown and Hagoort, 1993; e.g., Deacon et al., 2000). As presented above, some priming effects (e.g., orthography/phonology and morphology) systematically arise at shorter PPTs, while others (e.g., semantics) typically arise at longer PPTs. Manipulating PPTs thus allows us to test whether semantic, morphological and orthographic priming are "on line" at different points in processing and whether these effects are maintained in aging.

The facilitation effect of morphologically-related primes on target-word RTs has been more recently demonstrated in French-speaking participants. Quémart et al. (2011) compared young adult and child participant groups with forward-masked priming using different PPTs (60, 250 and 800 ms[1]) and four types of prime: pseudo-morphemic (such as *baguette—bague* 'baguette—ring,' where there is no shared stem between the prime and the target but a pseudo-parse is possible *bague* + *ette* 'ring + diminutive'), morphologically derived forms (*tablette—table*, 'shelf—table'), orthographic (*abricot—abri* 'apricot—shelter') and semantic (*tulipe—fleur*, 'tulip—flower'). Their results are globally similar to those of Feldman and Prostko (2002) in that morphological priming effects were constant across PPTs and groups, while other effects (orthographic and semantic) varied by group and PPT. At 60 ms PPTs, orthographic and semantic effects were non-significant in children and adults, while pseudo-morphological forms showed similar priming to true morphological priming: responses were faster by 25 ms on average. In adults, when the PPT was increased to 250 ms the data converged on previous results, with significant morphological priming (56 ms), no semantic, or orthographic priming, and a disappearance of pseudo-derived priming effects. The authors interpreted this as signaling an early morpho-orthographic processing stage, followed by a later lexical-semantic stage, linked to longer PPTs and prime perceptibility (Meunier and Longtin, 2007; Quémart et al., 2011). Similarly, an event-related potential (ERP) study by Royle et al. (2012) tested inflectional morphology, semantic, and orthographic priming effects on inflected verb recognition in French-speaking young adults, using sandwich-masked priming and a PPT of 50 ms. No significant RT differences were found for orthographic or semantic priming relative to unrelated pairs, while morphological priming reliably sped up target recognition by 29 ms on average. Additional ERP data from this study support autonomous morphological processing, as the authors found strong and long-lasting ERP modulation for morphological priming, weak and short-lasting modulation for orthographic priming, and no effects for semantic priming.

Globally, these results point to the interpretation that French speakers with ages varying from beginning readers to young adults rely on morphemic information when processing words, and that this process is distinct from orthography and semantics. Importantly, masked priming is known to block semantic effects, therefore allowing us to test whether morphological effects are parasitic on semantic ones, as suggested by eliminativists. The ERP study did not manipulate PPT, and it could be argued that at longer PPTs semantic processes could play a role in morphological priming. The present study therefore addresses this issue using the same paradigm and virtually the same

[1]We report results for PPTs used with adults—60 and 250 ms—but additional interesting results are reported for children at 800 ms, we refer the interested reader to this article for more details.

materials as Royle et al. (2012) in young and older adults, while manipulating PPT as in Feldman and Prostko (2002).

Aging and Cognition

A motivation for our age group comparison is that it is unresolved whether morphological representation and processing changes or remains stable at later stages of the lifespan. Aging is a multidimensional process that produces changes, to various extents, in different brain and cognitive functions. Older adults, compared to younger ones, show structural changes in terms of gray and white matter volume reduction (Raz et al., 1997, 2004; Good et al., 2001; Fjell et al., 2013; Lindemer et al., 2017) or reduced brain connectivity (Montembeault et al., 2012). These age-related changes are often accompanied by performance decline in additional cognitive processes such as executive function (i.e., attentional control, inhibition, working memory, and task monitoring), and episodic memory (Gordon and Kurczek, 2014). Word processing is thought to remain relatively stable in older adults. However, studies comparing younger and older adults have provided conflicting results depending on the nature of the task employed. On the one hand, it has been shown that older adults exhibit lower performance than younger ones in lexical retrieval tasks using explicit and effortful paradigms such as picture naming and word naming from definitions (Bowles and Poon, 1985; Verhaegen and Poncelet, 2013). However, it must be noted that these tasks require word retrieval or word planning. Both these processes rely on executive function abilities (such as attention) (Murray, 2000; Roelofs and Piai, 2011), which are sensitive to aging (Sylvain-Roy et al., 2015). On the other hand, when executive function effects are minimized using tasks that do not require explicit lexical access, such as priming paradigms with word stimuli, studies have often reported comparable performance in older, and younger participants (e.g., Lustig and Buckner, 2004). Together, these findings seem to suggest that (1) word representations are preserved in older populations and, (2) priming tasks could clarify the role of morphological representation in lexical access while minimizing the impact of other cognitive processes. For example, previous priming studies comparing older and younger adults have shown that semantic and orthographic priming are preserved in older adults (Madden et al., 1993).

Although aging adults show lexical access difficulties (e.g., tip-of-the-tongue states), few differences are observed between older and younger adults in language processing tasks, but inflectional morphology is surprisingly neglected. Most studies have investigated semantic priming effects (see Giffard et al., 2003 for a review) and very few (see Reifegerste et al., 2018) have investigated inflectional morphology priming. We present some studies of lexical access involving semantic, orthographic, and compound morpheme priming before presenting studies of inflection priming proper. Older participants show equal or stronger semantic priming than younger ones (see Burke, 1997, for a review), and evidence for deterioration in confrontation naming varies depending on studies, can be subtle, and may appear only in the seventh or eighth decade of life (Feyereisen, 1997; Connor et al., 2004; Zec et al., 2005; Obler et al., 2010) thus supporting robust lexical representation despite declining cognitive abilities. However, some have argued that older speakers have a higher dependence on whole word or semantic processing with aging (Patterson et al., 2007; Grieder et al., 2012; Provost et al., 2016; Chapleau et al., 2017). (Grieder et al., 2012) show similar strengths in aged (60–78 years) and younger Swedish-speaking adults, using semantic priming tasks in conjunction with ERP methods. Some studies show phonological difficulties in word production (MacKay and James, 2004) and priming, suggesting this might be a domain of weakness (James and Burke, 2000). Moscoso Del Prado Martín (2017) observes a decrease in morphological diversity in spontaneous speech with aging. No aging effects are found for the magnitude of morphological priming in Spanish compounds (e.g., *pasa—PASATIEMPO* 'pass—pastime') (Duñabeitia et al., 2009). Thus, morphological processing has not yet been shown to be a clear domain of weakness. Note that even though the reviewed studies mostly focus on lexical-semantic or phonological-orthographic processing abilities, these are relevant to models of morphological processing, especially those that argue that morphological processing is an epiphenomenon of semantic and orthographic processing.

Inflectional Morphology and Aging

A small number of recent studies using priming paradigms have focused more specifically on inflection processing in aging German speakers. These studies generally assume that morphological processing is available to all speakers, and investigate whether different types of inflection (namely regular vs. irregular, the first being more likely to be morphologically parsed) are processed in similar or different ways. Using cross-modal (auditory-visual) unmasked priming with no interval between the prime and target, Reifegerste and Clahsen (2017) establish that German-speaking adults (aged 50–83, $N = 32$, 23 women) show strong inflectional morphologial priming effects for regular inflected adjectives (e.g., *blaue—blaues* 'blue'), but weaker priming for irregular verbs (e.g., *werfen–wirft*, 'to-throw–throws'), as compared to identity priming (e.g., *wirft—wirft*). In a second study Clahsen and Reifegerste (2017) compare priming for regular and irregular verbs with similar participant groups and methods. They show less priming for irregular (e.g., *geschlafent—schlafe* 'slept–sleeps') vs. regular verbs (e.g., *getanzt–tanze* 'danced–dances') and, contrary to younger adults, older German speakers do not show priming for irregular verbs. Finally, Reifegerste et al. (2018) observe similar magnitudes of priming, in older and younger German speakers, for forward-masked derivation priming (e.g., *Warnung–warnen* 'warning–to-warn') and regular inflection priming (vs. *gewarnt–warnen* 'warned–to-warn'). No reliable orthographic (e.g., *Kasse–Kasten* 'cash register–box') or semantic priming (e.g., *Tisch–Stuhl* 'table–chair') was found. This body of work suggests that inflection morphological processing is stable in older adults but that some aspects of irregular inflection processing, possibly linked to the long-term memory storage of lexical representations (Reifegerste et al., 2018) or orthographic/phonological processing, can be less efficient. A few issues remain. No direct comparison to younger adults is made in Reifegerste and Clahsen (2017), and they compare regular adjectives to irregular verbs. The grammatical information on verb tense/person/number inflection vs. case/number marking on adjectives is a potential confound making their conditions less than ideal for direct comparisons. The second study (Clahsen and Reifegerste, 2017) contains no semantic, nor orthographic, control conditions to test for meaning or form overlap between prime and target, and, astonishingly, repetition priming did not show any advantage over unrelated priming[2]. Reifegerste et al. (2018) do have orthographic and semantic control conditions, but use different target words (see examples above), which is not ideal.

Thus, the study of inflectional morphological priming in young and older adults appears to provide critical information allowing us to better understand the effects of aging on word morphological representation. Our study will improve the state of knowledge on inflection morphological processing in aging as we will tease apart aging effects on semantic, orthographic, and morphological representations on lexical access. Based on results of previous studies, we hypothesized that young adults would show robust and equivalent morphological priming effects (i.e., faster RTs) across different PPTs, while orthographic effects would be absent or inhibitory in long PPTs (66 and 150 ms) but present when the prime is not perceptible (33 ms PPT). If at all, semantic priming was expected only when the prime was robustly perceptible (150 ms PPT). We expected older adults to show globally similar patterns as younger ones since, as far as the literature shows, no impairment specific to regular inflectional morphological processing has been supported, excluding MacKay and James (2004) and Moscoso Del Prado Martín (2017).

METHODS

Participants

Twenty-four young adults (aged 19–28, 12 women, two left-handers one for each sex) and 25 older adults (aged 61–80, 12 women, one male left-hander) participated in the study. Eleven older adults were between ages 61 and 70, and 14 between 71 and 80. No participant had a history of language or reading impairment, neurological damage or impairment. They all had normal or corrected-to-normal vision. All were native French speakers and lived in predominantly French-speaking environments. In order to exclude the presence of mild cognitive impairment, all participants were screened using the Mini-mental (French adaptation of the MMS, Folstein et al., 1975; Derouesné et al., 1999). Older participants were additionally evaluated using the Montreal Cognitive Assessment (MoCA, Nasreddine et al., 2005). An X-test was also used to evaluate basic motor response times. This was used to obtain a measure of simple visual reaction times. In this test, participants were presented with a white fixation dot in the center of a black screen of the laptop, followed after variable time intervals by a white cross target stimulus. Participants were asked to press the space bar on the laptop keyboard as quickly as possible when the white cross appeared on the screen. Scores on the MMS did not vary according to participant group (Younger: $M = 29.23$, $SD = 0.75$; Older: $M = 29.08$, $SD = 0.86$, t-test assuming equal variance = 0.62 $p = 0.269$). No differences were found on MMS scores between the older sub-groups (61–70: $M = 29.09$, $SD = 0.70$; 71–80, $M = 29.07$, $SD = 0.99$, $t = 0.5$, $p = 0.48$). MoCA scores in older adults were 28.21 on average ($SD = 1.25$). No difference was observed on the MoCA between participants aged 61–70 ($M = 28$, $SD = 1.18$) and participants aged 71 to 80 ($M = 28.38$, $SD = 1.33$, $t = -0.25$, $p = 0.40$). Participants did not differ on education levels (Younger: $M = 15.64$, $SD = 1.71$, range 14–22; Older: $M = 16.12$, $SD = 3.18$, range 13–25, t-test assuming unequal variance = −0.64 $p = 0.262$). However, both groups differed on their mean motor responses to the X-test (Younger: 291.65 ms, SD = 38.56; Older: 330.67, $SD = 42.67$, t-test assuming equal variance = −3.27 $p = 0.002$, see **Supplementary Figure 1**). Participants received 40$ CAD for their participation and signed informed consent to participate. The project was approved by the ethics committees of the *Université de Montréal* Faculty of Medicine, and the *Center de recherche de l'Institut universitaire de gériatrie de Montréal* (CRIUGM).

Procedure

After reading and signing informed consent, younger participants were comfortably seated in a quiet room at the *Université de Montréal*, and older participants at the CRIUGM, in front of a *HP D8907 p720* computer screen 45 cm from their face. They were asked to judge whether letter strings presented on the screen were real words of French or not. They were asked to respond as rapidly and accurately as possible. They were advised that the string would be preceded by a series of hash marks. They were not explicitly told about the prime, just that they had to respond after the hash marks. The participants' dominant hand was used to respond YES, and the non-dominant hand for NO, using S and L keys. Stimuli were presented with *Eprime 2.0* (Zuccolotto et al., 2012) and response accuracies (RAs) as well as reaction times (RTs) were recorded with this program. Each trial started with the presentation of a fixation point (+), a forward mask (########) for 500 ms, followed by a prime that was replaced after 33, 66 or 150 ms, by a backward mask for 20 ms, and then by the target on which participants made a lexical decision. The target disappeared when a decision was made, or else after 2,500 ms. A white screen followed for 1 s, and a new trial started. Two pauses were programmed within each list. Participants controlled pause length. No feedback was given on responses, and total testing time lasted between 45 and 55 minutes.

Stimuli

Here we present a brief description of priming conditions, timing and list development, see Royle et al. (2012) for details on stimulus properties. A master list using 42 regular target verbs and their six priming conditions was developed. The morphological, orthographic and semantic conditions

[2]They report confusion in some participants as to the task instructions, and specifically when confronted with repetition priming. These participants should probably have been discarded from the study.

each had their control prime matched for frequency and length. The semantic priming condition involved synonyms or troponyms (e.g., *brise—CASSE*, 'break—break') to mirror the semantic relationship between inflected verb forms as closely as possible. The orthographic condition involved a prime that was orthographically similar to the target without shared morphological structure or semantics (e.g., *cassis—CASSE,* 'blackcurrant—break'). In the morphological condition, the third person singular imperfect form of the verb was used as a prime (*cassait—CASSE*, 'broke—break'). Orthographic and morphological pairs were individually matched on the amount of orthographic, phonological and syllabic overlap they shared with the target. All primes were matched on form frequency and length (letters, phonemes and syllables), and in the case of morphological controls, internal word structure (e.g., *disait-CASSE,* 'said-break' matched with *cassait-CASSE*). We used the same target in all conditions in order to be able to directly compare different priming effects on target recognition. The prime and target were always presented in different fonts to avoid retinal imprints and visual overlap on the presented items. Forward and backward masks were used to reduce conscious priming and erase retinal imprints. The three PPTs used (33, 66, and 150 ms) were repeated across conditions (i.e., 42 items in 6 priming conditions and 3 PPT times). Based on this master list, nine lists were created with pseudo randomized orders, making sure that each target was not seen more than twice in each list (always in a given condition, with a specific PPT and its control priming condition). Within a list, 84 prime-target pairs were distributed into four blocks such that the same target was presented in maximally distant blocks of items (i.e., blocks 1–3 or 2–4). In younger adults, 42 real word filler pairs, 126 non-word target filler items and 8 practice items were added to each list. In total, each list contained 252 items and took no more than 15 minutes to complete. The stimuli used in older adults were the same as for younger ones except for 9 items that were excluded from the analyses in younger adults (see Statistical analyses below). Thus, older adults were presented with 33 critical items in each of the 6 experimental conditions, at three PPTs, randomized across nine lists, for a total of 234 items per list, including fillers. All participants were tested by the third author.

Testing Sessions

Participants came to three or four different 50-minutes sessions in the lab over a 3- or 4-week period: young adults were presented with three lexical decision stimulus lists per visit, while older adults were presented with two on the first three visits and three on the last one. These were presented in pseudo-random orders to each participant in order to avoid biases linked to sequence effects. Thus, all participants saw all nine lists. During the first meeting, all participants filled in a demographic questionnaire with specifics on their communication habits and medical antecedents. This was followed by their first experimental list, a visual acuity test, the MoCA cognitive screening test and the following list(s). In the second session, between the two first experimental lists, they were given the Mini-mental cognitive evaluation. During the third session, participants were tested on the X-test in addition to their lists. An average response time score was calculated for each participant. Finally, on the fourth meeting, all older participants completed three lexical decision stimulus lists.

Statistical Analyses

Following preliminary analyses of the young adult data, two participants were excluded because they had more than 20% error rates. Each item's recognition score was calculated and those with a level below 75% were eliminated from the analysis (the nine items were: *bride* 'put a bridle on,' *chipe* 'steal,' *fane* 'wilt,' *farde* 'put on makeup,' *ferre* 'shoe a horse/nail a horseshoe,' *fuse* 'burst forth,' *hume* 'inhale/smell,' *hale* 'haul,' and *larde* 'pierce/hurt'). Thus, 22 young participants (aged 19–28) and 33 targets were retained. Older adults were tested only on this subset in items and analyses were run on this subset in both groups. A response contingent RT analysis was performed on critical items (9.74 % of the data were excluded). Extreme responses beyond 1.5 s, were also eliminated (this accounted for 0.57 % of the data).

Target accuracy and reaction time data were subjected to linear mixed effect models using the lme4 package 1.1.12 in R (Bates et al., 2015). Fixed effects included in models were PrimeType (morphological, orthographic semantic,—minus their control conditions in RT data, see below—dummy coded with morphological prime as reference level, as it is at the core of the present study), prime presentation time (zPPT, as a standardized version of the continuous variable, PPT, using the rescale function from the *arm* R package, Gelman and Su, 2018), and zAgeGroup (a standardized version of the factorial binary variable AgeGroup). zPPT and zAgeGroup were normalized given their two very different scales. This procedure centered these two variables on 0, allowing us to establish a meaningful intercept that is the mean priming effect (at the reference level of PrimeType) independent of age group and PPT. Random-effect factors Subject and Item, as well as random slopes for Trial (chronological trial order within a list, continuous, *z*-scored: zTrial), testing Day (continuous, centered: cDay), list presentation Order (continuous, centered: cOrder) and mean reaction time on the X-test (RTX) were used in target accuracy analyses. Only random intercepts with an *intraclass correlation coefficient* (ICC) of more than 0.05, calculated with the *sjstats* package (Lüdecke, 2017), were retained. In RT analyses, we

TABLE 1 | Glmer model for response accuracy: fixed effects for the factors Group and PrimeType.

	Estimate	Std. error	*t*	Pr (>|t|)
Intercept	2.5573	0.2312	11.061	<0.0001
Orthographic control	−0.0053	0.0899	−0.059	0.95298
Semantic control	−0.0236	0.0896	−0.263	0.79246
Morphological	0.2695	0.0949	2.837	0.00455
Orthographic	0.3668	0.0971	3.779	0.00016
Semantic	−0.0326	0.0894	−0.364	0.71549
Group (Older Adults)	1.4869	0.3132	4.747	<0.0001

Nb. of observations = 27834, log likelihood −5236.3. The reference level for group is younger adults and for prime is morphological control.

TABLE 2 | Response accuracy and response times (RTs) in means (and standard deviations) for all priming conditions in each participant GROUP (younger vs. older).

	Morphological		Orthographic		Semantic	
	Primed	**Control**	**Primed**	**Control**	**Primed**	**Control**
YOUNGER ADULTS						
Response accuracy	91.8 (27.4)	89.4 (30.8)	92.8 (25.8)	89.9 (30.1)	89.5 (30.6)	89.4 (30.8)
Reaction time	575.0 (160.7)	608.0 (168.6)	596.1 (165.9)	619.0 (167.6)	607.5 (164.1)	615.5 (170.4)
OLDER ADULTS						
Response accuracy	98.1 (13.8)	97.7 (14.8)	97.9 (14.2)	97.2 (16.4)	97.3 (16.2)	97.5 (15.5)
Reaction time	729.3 (170.6)	749.9 (171.2)	740.7 (171.5)	756.5 (170.9)	746.4 (166.5)	752.5 (168.0)
OVERALL						
Response accuracy	95.1 (21.5)	93.8 (24.0)	95.5 (20.6)	93.8 (24.1)	93.7 (24.4)	93.7 (24.0)
Reaction time	659.5 (183.1)	685.6 (184.1)	674.9 (183.7)	694.8 (182.7)	684.1 (179.3)	691.1 (182.3)

calculated differences between control and primed conditions. This precluded the use of TRIAL as a random-effect factor for these analyses. We started with a maximal random-effects structure, simplifying the model in cases of convergence failure, and modeled factors as interactions until we reached the best fit, determined by comparing two minimally different models using the anova() function to perform the likelihood ratio test, and assess the significance of effects of the fixed factors as main effects as well as their interactions. This was assessed by comparing a model containing the interaction in question, either to a model containing only the relevant lower-level interactions or to a model containing only the relevant main effects. In order to make the selected models easier to interpret, we used the ANOVA wrapper (Type III Wald chi-square test) in the *car* package (Fox and Weisberg, 2011). When needed, *post-hoc* pairwise comparisons were performed using the *emmeans* package 1.3.0 in R (Lenth, 2018). A first analysis involved response accuracy (RA) and a second one modeled response-contingent reaction times (RTs).

RESULTS

Response Accuracy

We tested for non-linear effects of covariates and for a contribution of by-participant random slopes for zTRIAL, cDAY, RTX, and cORDER. We established these did not significantly improve the model, nor did the random factor ITEM (ICC = 0.0417). SUBJECT was included as a random factor (ICC = 0.0846). The maximal structure containing all fixed effects (GROUP, PRIMETYPE, cPPT, cAGE) and the random factor SUBJECT failed to converge. The best model for RAs included the fixed effects GROUP and PRIMETYPE but no interaction of these (AIC: 10488.5, see **Tables 1**,**2**). PPT never reached significance as a simple effect nor in interaction with other simple effects. *Post-hoc* contrasts using *lsmeans* revealed significant differences between orthographic control and orthographic priming conditions, $t = -3.83$, $p = 0.0017$, morphological, and semantic priming conditions, $t = 3.197$, $p = 0.0174$, as well as orthographic and semantic priming conditions, $t = 4.135$, $p = 0.0005$. In other words, we found small but significant priming effects for orthography and morphology (i.e., higher accuracy for primed compared to control items), whereas no such effect was observed for the semantic condition. GROUP effects reflected the fact that older adults showed higher levels of correct responses (97.64%, $SD = 0.15$) than younger adults (90.46%, $SD = 0.29$), who in turn displayed more within-group variability (as reflected by SDs). In essence, scores are globally high in all groups across all conditions, especially in older adults, with the lowest scores found in control and semantic priming conditions.

Reaction Time Data

Older participants were on average 102 ms slower than young adults, which is some 62 ms more than the motor-related differences found in the X-test (see **Table 2** for RTs by condition). In order to quantify priming effects, we subtracted each primed-condition reaction time from their control-condition reaction time for each item, at each PPT, within each participant. These data, summarized in **Table 3**, were then entered into mixed models. We tested for non-linear effects of covariates and for a contribution of by-participant random slopes for cDAY and cORDER (TRIAL could not be measured due to the subtraction process, while RTX effects did not significantly contribute to the model). Surprisingly, adding the factor zPPT to the model did not improve its likelihood either, but this factor was maintained in order to illustrate a lack of effect, as it was part of the experimental design. Models were fit using the maximal random effect structure justified by the design that would converge on the data (Barr et al., 2013). The maximal structure contained all fixed effects (zAGEGROUP, PRIMETYPE, zPPT), their interactions, by-SUBJECT and by-ITEM random intercepts as well as by-ITEM random slopes for the interaction PRIMETYPE*zPPT (AIC:

TABLE 3 | Average priming effects for Groups, Prime type and PPT (non-weighted means).

Group	Prime type	PPT 33	PPT 66	PPT 150	Average
Young	Morphological	37.06	39.36	27.72	34.71
	Orthographic	25.38	19.63	26.05	23.69
	Semantic	15.32	12.74	1.28	9.78
Older	Morphological	16.08	19.12	22.82	19.34
	Orthographic	17.33	22.05	11.50	16.96
	Semantic	11.20	10.67	−0.16	7.24
Average	Morphological	26.57	29.24	25.27	27.03
	Orthographic	21.35	20.84	18.77	20.32
	Semantic	13.26	11.71	0.56	8.51

TABLE 4 | Lmer model for priming effects with fixed effects of PrimeType, zAgeGroup, and zPPT as well as the interaction between PrimeType and zAgeGroup.

	Estimate	Std. error	*t*	Pr (> |t|)
Intercept	26.004	6.692	3.886	<0.001
Orthographic prime	−5.039	10.982	−0.459	0.649
Semantic prime	−17.944	6.117	−2.933	0.006
zAgeGroup	−15.443	5.583	−2.766	0.006
zPPT	−5.138	3.902	−1.317	0.197
Orthographic prime:zAgeGroup	8.234	7.575	1.087	0.277
Semantic prime:zAgeGroup	13.030	7.610	1.712	0.087

As random effects, random intercepts for Subject and Item are included, as well as by-Item random slopes for the effects of PrimeType and zPPT. The reference level for PrimeType is morphological prime. Nb. of observations = 12,550.

165010.8). This model included the fixed effects PrimeType, zAgeGroup, zPPT and the interaction between PrimeType and zAgeGroup, by-Subject and by-Item random intercepts as well as by-Item random slopes for PrimeType and zPPT effects (AIC = 165034.8). This model does not significantly differ from the maximal model ($\chi^2_{(1)} = 24.02$, $p = 0.09$). Results of the corresponding analysis are presented in **Table 4**, and **Figure 1** illustrates the underlying data by PrimeType and zAgeGroup (but collapsed across PPT levels, as PPT never contributed significantly to the models).

The data can be summarized by three main observations (**Figure 1**). First, all priming effects were numerically larger in the group of young adults. Secondly, in both age groups, morphological priming showed the strongest RT effect, followed by orthographic priming, and finally by a very small and non-significant semantic priming effect. Thirdly, the group differences were most prominent in the morphological condition. Thus, whereas young adults' response times seemed to profit considerably more from morphological priming (35 ms) than from orthographic priming (24 ms), in older participants both types of priming reduced response times to almost the same extent (19 and 17 ms, respectively). These observations are largely confirmed by statistical analyses, which however also pointed to some ambiguities.

With morphological priming being the reference level for PrimeType, and age groups being represented by a normalized and centered variable (zAgeGroup), all effects in **Table 4** must be interpreted relative to morphological priming across both age groups. The overall morphological priming effect (intercept of 26 ms) was highly significant and did not statistically differ from the orthographic priming effect (21 ms), while a significant difference was found between morphological and semantic priming (8 ms), suggesting that the latter was not significantly different from zero. zPPT did not contribute to any significant effects. Crucially, morphological priming was significantly modulated by zAgeGroup ($p < 0.006$), i.e., the priming effect of 35 ms in young adults was significantly larger than the 24 ms effect in older adults. On the other hand, the difference between morphological and semantic priming was only marginally influenced by zAgeGroup ($p < 0.09$). Non-significant results for the difference between orthographic and morphological priming ($p = 0.649$) and for its modulation by age (i.e., Orthographic prime:zAgeGroup; $p = 0.277$) indicated that orthographic priming showed a similar pattern as morphological priming. Thus, based on this analysis alone, the RT data seem to suggest that (a) orthographic and morphological conditions both showed comparable (significant) priming effects in both groups (although more so in young participants), and that (b) they equally differed from the semantic condition that did not show any priming. However, these assumptions may be an oversimplification. An additional direct contrast between orthographic and semantic conditions revealed that these two conditions did not statistically differ from each other either ($p = 0.4219$). In fact, when running the same model as above, but with the *semantic* priming condition as the reference, the results suggest a distinct pattern, as shown in **Table 5**. Now, the orthographic priming condition seems to pattern with the semantic rather the morphological condition, i.e., neither the orthographic priming effect nor its interaction with zAgeGroup differ from the (non-significant) patterns of the semantic condition. The only significant effect in this version of the model is (again) the difference between semantic and morphological priming ($p = 0.006$), which moreover marginally interacted with age ($p = 0.087$). In other words, whereas a clear morphological priming effect could be reliably distinguished from a very weak (or absent) semantic priming effect *irrespective of the model version*, the status of the orthographic priming condition remained unclear as it did not significantly differ from either morphological priming (**Table 4**) or from semantic priming (**Table 5**). Intriguingly, the ambiguity regarding orthographic priming also extends to the important question of whether or not its magnitude differed between young and older participants. **Table 4** suggests it is modulated by age—as with morphological priming—whereas **Table 5** suggests it is not modulated by age—similar to the semantic priming condition. Moreover, the finding that, in both models, the impact of age groups on the difference between semantic and morphological priming was only marginal ($p = 0.087$) results in a number of possible interpretations, depending on whether a marginally significant interaction is treated (a) as a non-significant result or rather (b) as a suggestive

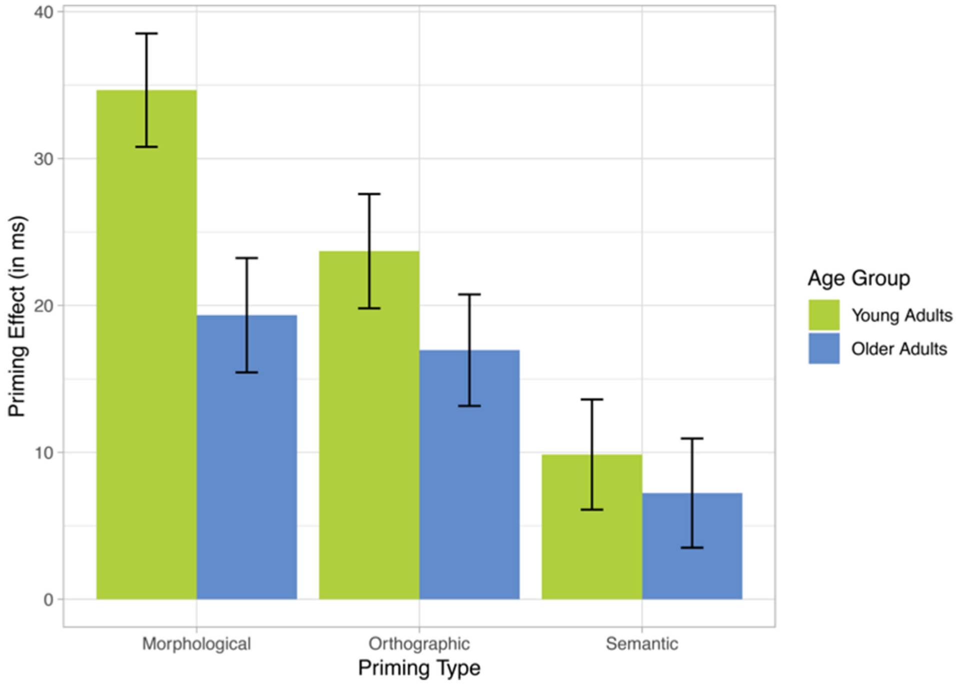


FIGURE 1 | mean primary effects (RTs in control minus primary condition), and standard errors, in morphological, orthographic, and semantic conditions, in younger and older French-speaking adults.

TABLE 5 | Lmer model for priming effects with fixed effects of PrimeType, zAgeGroup and zPPT as well as the interaction between PrimeType and zAgeGroup.

	Estimate	Std. error	*t*	Pr (> |t|)
Intercept	8.060	6.025	1.338	0.190
Morphological prime	17.944	6.117	2.933	0.006
Orthographic prime	12.905	10.517	1.227	0.229
zAgeGroup	−2.412	5.612	−0.430	0.667
zPPT	−5.138	3.902	−1.317	0.198
Morphological prime:zAgeGroup	−13.030	7.610	−1.712	0.087
Orthographic prime:zAgeGroup	−4.797	7.597	−0.631	0.5277

As random effects, random intercepts for Subject and Item are included, as well as by-Item random slopes for the effects of PrimeType and zPPT. The reference level for PrimeType is semantic prime. Nb. of observations = 12,550.

effect. In the former case (a), **Table 4** with its highly significant zAgeGroup effect for morphological priming and no further interactions involving age groups suggests (i) an age effect being *present across the board* in all three priming conditions; whereas **Table 5** suggests (ii) that age effects were *absent across the board*. In the latter case (b), the pattern would suggest either (iii) that age influences were *present only for morphological and orthographic conditions*, but not for semantics (**Table 4**), or (iv) that *age affected only the morphological priming condition* (**Table 5**).

In our opinion, the inconsistencies described above point to a rather ambiguous pattern of results that call for additional analyses. To further clarify the actual pattern, we ran separate models within each priming condition—and found three different patterns.

Separate Models for the Three Priming Conditions and for Morphological Priming vs. Morphological Control Conditions

We developed models for each level of PrimeType separately while again investigating whether zPPT and zAgeGroup would affect these. As zPPT never led to any significant effects (all p's> 0.16), we will not report these data separately.

A first model for morphological priming (**Table 6**) revealed both a highly significant priming effect ($t = 3.858$; $p < 0.001$) and a strong main effect of zAgeGroup ($t = -2.609$; $p = 0.0124$), confirming that morphological priming was significantly reduced in older participants. To better understand if the group of older participants showed any evidence of morphological priming, we followed up with analyses in each group. These *post-hoc* analyses demonstrated that morphological priming was very strong and extremely significant in young adults ($\beta = 34.2$ ms; $p = 2.2e\text{-}16$; based on 1,848 observations), whereas it was considerably weaker—though still highly significant—in older adults ($\beta = 19.2$ ms; $p = 7.158e\text{-}07$; based on 2,342 observations). Note that this result is exactly what we found when running models on original unsubtracted RT data for morphological vs. control priming. In that analysis the group differences were reflected by a significant morphological priming by age group interaction ($p = 0.02$).

TABLE 6 | Lmer model for morphological priming with fixed effects of zAgeGroup and zPPT, as well as their interaction.

	Estimate	Std. error	*t*	Pr (> |t|)
Morph Intercept	26.098	6.764	3.858	<0.001
Morph zAgeGroup	−15.568	5.968	−2.609	0.012
zPPT	−1.281	6.384	−0.201	0.842
zAgeGroup:zPPT	15.078	10.891	1.384	0.166

As random effects, random intercepts for Subject and Item are included, as well as by-Item random slopes for the effect of zPPT. Nb. of observations = 4,190.

TABLE 7 | Lmer model for orthographic priming with fixed effects of zAgeGroup and zPPT, as well as their interaction.

	Estimate	Std. error	*t*	Pr (> |t|)
Orth Intercept	20.964	6.584	3.184	0.003
Orth zAgeGroup	−7.156	5.376	−1.331	0.183
zPPT	−3.360	7.040	−0.477	0.637
zAgeGroup:zPPT	−8.050	10.746	−0.749	0.454

As random effects, random intercepts for Subject and Item are included, as well as by-Item random slopes for the effect of zPPT. Nb. of observations = 4,213.

TABLE 8 | Lmer model for semantic priming with fixed effects of zAgeGroup and zPPT, as well as their interaction.

	Estimate	Std. error	*t*	Pr (> |t|)
Sem Intercept	8.193	5.970	1.372	0.180
Sem zAgeGroup	−2.376	5.252	−0.452	0.651
zPPT	−11.296	7.153	−1.579	0.124
zAgeGroup:zPPT	2.756	10.506	0.262	0.793

As random effects, random intercepts for Subject and Item are included, as well as by-Item random slopes for the effect of zPPT. Nb. of observations = 4,147.

The second model for orthographic priming (**Table 7**) revealed an average priming effect of 20.96 ms, which was significantly different from 0 ($t = 3.184$; $p = 0.003$). Unlike morphological priming, orthographic priming was not affected by zAgeGroup ($p > 0.18$). Lastly, a model for semantic priming (**Table 8**) showed that the priming effect of 8.193 ms was not significantly different from 0 ($p > 0.18$) and was not modulated by age ($p > 0.65$). Again, this result is exactly what we found when running models on original RT data for orthographic and semantic priming vs. their controls. In contrast to the morphological priming comparison, no main effects for orthography or semantics nor significant priming by age group interactions were observed (p's > 0.2).

DISCUSSION

Our study extends priming paradigms to a less studied language (French) and to a population of healthy aging adults, in order to evaluate the effect of aging on morphological, orthographic, and semantic representation of words. Our experimental design differs from previous approaches in several important ways. First, we used the exact same target words in all six priming conditions and meticulously matched the primes on psycholinguistic properties (e.g., in terms of orthographic prime-target overlap between orthographic and morphological conditions). Secondly, we used inflectional rather than derivational morphology conditions—thereby better controlling for semantic overlap—as well as synonyms or troponyms as semantic primes. Third, we employed a "sandwich masking" technique that is known to suppress semantic priming (at least at short PPTs), allowing us to focus on the nature of morphological priming and its relationship to orthographic priming in the absence of semantic contributions. Fourth, we manipulated PPT at three levels (33, 66, and 150 ms). To our knowledge, this is the first study of healthy aging to investigate inflectional morphological processing with this type of design. Overall, our findings suggest that inflectional morphological priming can facilitate lexical processing in older adults, but that this effect is not the same as that observed in younger adults, as it seems to have weakened to the extent that it is numerically indistinguishable from orthographic priming. As we will discuss below, this result reaches beyond the area of cognitive aging, because it has general implications for morphological processing in psycholinguistic research.

General Aging Effects on Performance

Our data provide support for the hypothesis that, while aging slows lexical decision responses, it does not negatively affect accuracy (Lima et al., 1991; Madden, 1992; Myerson et al., 1992; Cohen-Shikora and Balota, 2016; Robert and Rico Duarte, 2016; Reifegerste et al., 2018). In fact, accuracy rates on the task were significantly higher in older than younger adults. Furthermore, younger adults showed more variability than older participants, with larger standard deviations for target recognition. These findings are consistent with previous studies showing that older adults perform at ceiling in lexical decision tasks with response accuracies comparable with, or even better than, younger adults (Lima et al., 1991; Madden, 1992; Myerson et al., 1992; Cohen-Shikora and Balota, 2016; Robert and Rico Duarte, 2016). On the other hand, and consistent with previous reports, older adults were generally slower than younger adults (by 102 ms on average in our study, also consistent with previous research, e.g., Ober and Shenaut, 2014; Curzietti et al., 2017). In order to determine if this significant difference was simply due to prolonged motor response execution (Falkenstein et al., 2006) or related to the specific psycholinguistic task, we also collected X-test data targeting pure motor planning and execution. We found that the aging-related slowing for lexical decision was above and beyond that of behavioral motor responses on the X-test, which were on average 42 ms slower in older adults as compared to younger participants. Assessing motor responses separately—as we did using the X-test—seems crucial to tease apart age effects on cognitive vs. motor behavior, and to avoid misinterpretations. Our data suggest that higher age increased the duration of cognitive processes involved in the lexical decision task by some 60 ms, i.e., aging appears to have affected cognition even more than motor control. A number of mechanisms accounting for these rather general (i.e., condition-independent) aging effects have been discussed in the literature. On the one hand, the frequent finding of longer processing times and higher performance accuracy in older adults has been described as a "time-accuracy tradeoff," potentially suggesting that older participants may be more cautious, and value correct responses more highly than fast ones (e.g., Ratcliff et al., 2000). In other words, longer response times in older participants—once corrected for rather trivial motor components—may reflect distinct processing *preferences* (accuracy over speed), not cognitive decline. However, this interpretation has not found consistent support in the literature (Kliegl et al., 1994; Myerson et al., 2003), which instead points to a rather complex pattern of (neuro-)cognitive aging effects (Hedden and Gabrieli, 2004). According to some authors, cognitive slowing in aging studies on lexical processing such as ours may reflect specific difficulties in orthographic stimuli processing, rather than general lexical access difficulties (Allen et al., 2002) or word representation deterioration. In fact, according to the "decision complexity advantage" hypothesis, older adults rely on progressively larger perceptual units in the attempt to compensate for their difficulties in encoding smaller orthographic units when processing visually presented words (Allen et al., 1993, 2002, 2011). However, given the lack of age effects on orthographic priming in our study, these differences in processing do not appear to be universal and may be linked to the specific conditions that Allen et al. tested (e.g., case manipulation within the word, e.g., TarGEt, or adding spaces between letters).

Lexical Priming and Aging Effects

Apart from the general age group differences discussed above, we observed significant modulations of accuracy levels in different priming conditions. However, differences were minimal between all conditions (2%, on average), and overall performance was near ceiling in older adults. Across age groups, both orthographic and morphological primes improved performance accuracy compared to unprimed controls, while no such effect was found for semantic primes. The lack of semantic priming was expected and, in fact, intended. Recall that we employed a masked priming technique known to suppress semantic priming, in order to test the "eliminativist" hypothesis that morphological priming can be reduced to a combination of semantic and orthographic priming (Bates and Goodman, 1997; Seidenberg and Gonnerman, 2000; Devlin et al., 2004). From this perspective, if semantic priming was successfully suppressed, morphological priming should be indistinguishable from orthographic priming. Accuracy data in both groups were clearly in line with this prediction. However, as in most priming studies with near-ceiling performance levels, the more revealing results were expected to come from RT data.

Surprisingly, even our longest PPT (150 ms) did not lead to significant semantic priming on RT data. Semantic priming has been reported in different priming paradigms using lexical processing (Balota and Duchek, 1988; Laver and Burk, 1993, for a meta-analysis see Feldman and Prostko, 2002; Giffard et al., 2003), and multimodal cross-modality priming (auditory-visual, Vallet et al., 2013). It must be pointed out, however, that the vast majority of these studies employed long SOAs (i.e., >200 ms) and no masking. Feldman and Prostko (2002) found semantic priming at PPTs of 116 and 300 ms, and so we expected to see a significant effect at a PPT of 150 ms. The absence of this effect (and of PPT modulations of priming effects in general) suggests that masked priming is more powerful in suppressing semantic priming than we originally anticipated. This was also confirmed by Quémart et al. (2011), who found semantic priming with prime masking only at a PPT of 800 ms (no adults were tested on this PPT). In hindsight, knowing now that, with sandwich masking, much longer PPTs are necessary for semantic priming, we believe using PPTs beyond 250 ms would have been necessary for it to emerge. It is conceivable that even with masking, semantic priming can occur when participants are asked to attend to the prime. In fact, semantic priming effects from behavioral and ERP studies have been argued by a number of researchers to be driven by factors such as directed attention to the prime, or strategic processing based on partial prime perception, e.g., letters (Abrams and Greenwald, 2000; Klinger et al., 2000; Kiefer and Brendel, 2006; Kouider and Dehaene, 2007; Kouider and Dupoux, 2007). In our study, participants performed on a lexical decision task, that is they were focused on the target and not the prime.

Response time data were found to be modulated by our prime types as well as age, resulting in a more complex pattern than that provided by our accuracy data. Specifically, morphological priming led to significantly reduced reaction times in both young (34 ms) and older (19 ms) adults, and the difference in magnitude of this effect between young and older adults (15 ms) was also found to be significant. As expected, there was no indication of semantic priming in either group, and the difference between (non-significant) semantic and (significant) morphological priming itself was significant across both groups.

For the orthographic priming condition, the results were less clear. Overall, the magnitude to which orthographic primes reduced response times on the target word (20 ms) was between the priming effects for morphology (27 ms) and those for semantics (8 ms). Depending on the statistical model, orthographic priming seemed to pattern either with the (significant) morphological priming effect (**Table 4**) or with the (non-significant) semantic priming effect (**Table 5**). Similarly, whether the difference in orthographic priming of 7 ms between young (24 ms) and older participants (17 ms) should be viewed as statistically significant (in line with the morphological condition) or as non-significant (in line with the semantic condition), partly depended on the reference condition in the model. An additional analysis focusing exclusively on the orthographic priming condition indicated that this priming effect (i) was clearly significant ($p = 0.003$), but (ii) did not differ between age groups ($p = 0.183$). In other words, whereas orthographic priming remains robust with increasing age, morphological priming—which at a young age is significantly stronger than orthographic priming—becomes weaker and virtually indistinguishable from orthographic priming.

Perhaps the most crucial question concerns the status of orthographic priming in relation to morphological priming. Recall that an eliminativist perspective would predict that, in the absence of any semantic priming, morphological priming should be entirely driven by orthographic priming. Since orthographic overlap between prime and target was meticulously matched between these two conditions, one would expect virtually identical priming effects. On the other hand, if the patterns for morphological priming can be shown to be different from orthographic priming, this would support the common assumption that morphology must be viewed as a distinct level of psycholinguistic representation and processing (Stockall and Marantz, 2006; Aronoff and Fudemann, 2011; Amenta and Crepaldi, 2012). Unfortunately, as discussed above, the various statistical analyses of our data did not warrant an unambiguous conclusion as to whether morphological priming effects in our study can be fully attributed to orthographic priming or should rather be viewed as distinct. However, we believe that taking other analyses and data into account ultimately favors the latter perspective. *First*, recall that the morphological priming condition resulted in a significantly larger effect in young compared to older adults, a group difference absent in the orthographic condition. This pattern is rather unexpected if we assume that both conditions rely on the exact same mechanism (i.e., orthographic priming), but would be expected if they differed qualitatively from each other. Additional evidence that the two conditions of our study involve distinct priming mechanisms (at least in young adults) comes from our previous ERP study (Royle et al., 2012), which used the exact same stimuli and the same "sandwich-masking" technique as the present one. In that study, even though morphological and orthographic priming conditions showed similar accuracy rates (in line with our present data), significant differences in the ERP patterns clearly demonstrated distinct underlying neurocognitive mechanisms. Importantly, the orthographic condition only showed ERP effects during an early time window (N250) known to reflect formal (orthographic/phonological) priming, whereas the morphological condition also affected subsequent lexical processing stages (reflected by the N400). The semantic condition did not show either behavioral or ERP priming effects (in line with the present study). Since both the population (young adults aged 18–35 years) and the PPT of some 50 ms were comparable to the present study, there is little reason to assume that the underlying cognitive mechanisms should differ between that study and our present group of young adults.

Secondly, if we assume that—in absence of any semantic priming—morphological priming is the same as orthographic priming and linked to orthographic feature overlap (the eliminativist view), any variability among either subjects or items should be comparable for the two conditions. Alternatively, if both rely on distinct mechanisms, no such relationship is expected. We pursued this logic by correlating priming effects across items and participants. Our initial set of correlations tested if those participants in our study who showed the strongest morphological priming were also the ones who showed the strongest orthographic priming. Neither within nor across groups did we find any significant effects supporting this hypothesis (all p values > 0.45). The second set of analyses ran the corresponding correlations across items, which was feasible because the same target words were used in all conditions. What we found was that items showing the strongest morphological priming were *not* those with the strongest orthographic priming, and *vice versa*. On the contrary, we observed a *negative* correlation between these two priming effects: $r^2 = -0.38$, $t_{(31)}$ -2.31, $p = 0.02$, two-tailed. These same negative correlations were found within younger adults when analyzed alone ($r^2 = 0.42$, $t_{(31)}$ -2.56, $p = 0.015$), while no significant correlation was observed in the older group ($r^2 = -0.23$, $t_{(31)}$ -1.34, $p > 0.1$). This is exactly the pattern one would expect, if (a) morphological priming is different from orthographic priming, and (b) morphological priming is stronger (or more prevalent) in young compared to older adults. Even though these results should not be taken as conclusive evidence for a distinct morphological priming mechanism, they are difficult to explain from an eliminativist perspective.

If we assume that morphological priming in young adults was indeed distinct from orthographic priming, a follow-up question concerns whether the priming effect we see in the morphological condition for *older* adults is still distinct from orthographic priming in that group. On the one hand, one could argue that morphological priming in young adults consists of orthographic priming plus "true" morphological priming. This interpretation appears to be in line with the ERP findings in Royle et al.'s (2012) study. If morphological priming effects in older adults are reduced to the levels of orthographic priming, then all that remains may be orthographic priming, i.e., "true" morphological priming is lost with aging. Alternatively, it is conceivable that, despite the same magnitude of effects, both types of priming are still qualitatively different. In this case, "true" morphological priming would be at least partly preserved. Our current analyses focusing on behavioral data alone are insufficient to distinguish between these two options. ERP data would certainly provide insight and complimentary data to accuracy and response-time data as to what the cognitive underpinnings for morphological and orthographic processing in aging are. For example, Royle et al. (2012) found similar accuracy for both orthographic and morphological priming but different ERP patterns, and Morgan-Short et al. (2012) demonstrated that explicit and implicit second-language learners relied on distinct neurocognitive mechanisms to process L2 grammar while behavioral patterns were indistinguishable. Applied to our present data, we can make clear predictions of what types of priming effects we would observe in neurocognitive ERP responses for older vs. younger adults. With masked priming and short PPTs, we expect semantic priming to have lesser or no effects on ERPs in older participants, similar to what has already been shown in younger adults. Orthographic priming should modulate the N250, as with younger adults. However, morphological priming could result in two different scenarios. Under the assumption that morphological priming is entirely carried by orthographic priming in older adults, only the N250 should be modulated by morphological priming. However, if we assume that morphological processing is still present although weaker in older adults, both the N250 and the N400 should be modulated by morphological priming, similar to young adults but to a lesser extent. Thus, behavioral and ERP approaches together provide complimentary information on language processing: while ERPs can tease apart the time-course of different processing stages, behavioral approaches allow for complex designs involving different prime-presentation manipulations, which in turn can help us define ideal manipulations for more constrained ERP experimental designs.

The facilitatory effect of morphological priming on lexical decision in our sample of young adults replicates Quémart et al's (2011) findings showing that that both French-speaking children and young adults take advantage of morphological priming. Similarly, Jacob et al. (2018) find that, in young adult German-speakers, inflectional and derivational morphology priming results in speeded reaction times as compared to a control condition, but that orthographic and semantic priming do not. This is also consistent with a recent ERP study on late-second-language learners demonstrating that morphological priming boosts word recognition in French learners, independent of learning level (Coughlin et al., 2019), and with previous ERP data showing significantly different modulation of early and late components for lexical processing with the same stimuli as ours, that is weak effects of orthographic priming, strong and longer-lasting effects of morphological priming and no semantic priming (Royle et al., 2012).

Our results critically expand previous research in the field of morphological processing and aging by demonstrating that regular-inflection morphological priming is lesser in older adults than younger ones, in contrast to what has been found in German by Reifegerste et al. (2018), who find equivalent regular inflection priming effects in younger and older participants. In line with the rare inflection morphology production studies (MacKay and James, 2004; Moscoso Del Prado Martín, 2017), we show that morphological processing might also be somewhat impaired in comprehension. However, our results do not provide evidence for formal (phonological/orthographic) difficulties in aging, contrary to MacKay and James (2004). We can reconcile the apparently contradictory data by proposing that morphological representation is available but weaker in older vs. younger adults. However, studies showing age effects on inflection morphology engaged older adults in complex language production tasks, which might be an additional source of difficulty. Based on our present data, the claim that morphological processing is not affected by aging (e.g., Reifegerste et al., 2018) might be an overgeneralization.

CONCLUSION

In conclusion, our study suggests that morphological word representation is not stable over the lifespan, and that this effect is independent from semantic abilities, but its relationship to orthographic processing remains unclear. Older adults showed regular inflection morphological priming at similar levels to orthographic priming, and no semantic priming. This indicates that inflectional morphological processing is possibly maintained and relatively efficient in the older population, and that whatever this effect is it is not an epiphenomenon of semantic processing, while orthographic processing is maintained over the lifespan. This is a crucial result that allows us to better define the nature of potential language processing difficulties in older adults. Because some studies show that morphological and phonological production in spoken language can be difficult in aging adults, the source of these difficulties has yet to be clearly identified. Future studies should focus on how morphological word representation interacts with other cognitive and language domains, in order to better understand morphological errors that are sometimes observed in speech production in older adults, for example. Complimentary ERP studies of aging might shed light neurocognitive mechanisms underlying morphological processing across the lifespan.

ETHICS STATEMENT

Participants received 40$ CAD for their participation and signed informed consent to participate. The Project was approved by the ethics committees of the *Université de Montréal, and the Centre de recherche de l'Institut Universitaire de Gériatrie de Montréal* (CRIUGM).

AUTHOR CONTRIBUTIONS

PR: study conception, statistical analyses, manuscript development, data interpretation. KS: study conception, data analysis, data interpretation. ÉD: participant recruitment, data collection, data management, research report, and preliminary analyses. ACH: statistical analyses. SMB: theoretical aspects of aging, experiment programming, participant recruitment, lab supervision, manuscript development. All authors participated in manuscript writing and revision.

FUNDING

This project was supported by Université de Montréal seed funding to PR, Subvention Institutionnelle du CRSH, UdeM (# R0014537), an SSHRC Insight grant (# 435-2013-2052) to KS and PR, a CRBLM Research incubator award (2012) to PR, SMB, Gonnerman, and KS, an SSHRC Partnership Grant to Libben et al (# 895-2016-1008), and two summer bursary undergraduate scholarships from the Faculty of Medicine of the University of Montreal (COPSE) (2013-2014) to ÉD.

ACKNOWLEDGMENTS

We thank participants who volunteered for this project and Cynthia Sarkis for help with data management.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2019.00016/full#supplementary-material

REFERENCES

Abrams, R. L., and Greenwald, A. G. (2000). Parts outweigh the whole (word) in unconscious analysis of meaning. *Psychol. Sci.* 11, 118–124. doi: 10.1111/1467-9280.00226

Alegre, M., and Gordon, P. (1999). Rule-based versus associative processes in derivational morphology. *Brain Lang.* 68, 347–354. doi: 10.1006/brln.1999.2066

Allen, P. A., Bucur, B., Grabbe, J., Work, T., and Madden, D. J. (2011). Influence of encoding difficulty, word frequency, and phonological regularity on age differences in word naming. *Exp. Aging Res.* 37, 261–292. doi: 10.1080/0361073X.2011.568805

Allen, P. A., Lien, M. C., Murphy, M. D., Sanders, R. E., Judge, K. S., and McCann, R. S. (2002). Age differences in overlapping-task performance: evidence for efficient parallel processing in older adults. *Psychol. Aging* 17, 505–519. doi: 10.1037//0882-7974.17.3.505

Allen, P. A., Madden, D. J., Weber, T. A., and Groth, K. E. (1993). Influence of age and processing stage on visual word recognition. *Psychol. Aging* 8, 274–282. doi: 10.1037//0882-7974.8.2.274

Amenta, S., and Crepaldi, D. (2012). Morphological processing as we know it: an analytical review of morphological effects in visual word identification. *Front. Psychol.* 3:232. doi: 10.3389/fpsyg.2012.00232

Ansaldo, J., Marsolais, Y., Methqal, I., Alary, F., and Joanette, Y. (2013). The adaptive aging brain: evidence from the preservation of communication abilities with age. *Eur. J. Neurosci.* 37, 1887–1895. doi: 10.1111/ejn.12252

Aronoff, M., and Fudemann, K. (2011). *What Is Morphology?* Oxford: Wiley-Blac.

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychol. Rev.* 118, 438. doi: 10.1037/a0023851

Balota, D. A., and Duchek, J. M. (1988). Age-related differences in lexical access, spreading activation, and simple pronunciation. *Psychol. Aging* 3, 84–93. doi: 10.1037/0882-7974.3.1.84

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bates, D., Mäechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Bates, E., and Goodman, J. C. (1997). On the inseparability of grammar and the lexicon: evidence from acquisition, aphasia and real-time processing. *Lang. Cogn. Process.* 12, 507–584. doi: 10.1080/016909697386628

Bertram, R., Laine, M., and Virkkala, M. M. (2001). The role of derivational morphology in vocabulary acquisition: get by with a little help from my morpheme friends. *Scand. J. Psychol.* 41, 287–296. doi: 10.1111/1467-9450.00201

Bijeljac-Babic, R., Biardeau, A., and Grainger, J. (1997). Masked orthographic priming in bilingual word recognition. *Mem. Cognit.* 25, 447–457. doi: 10.3758/BF03201121

Bowles, N. L., and Poon, L. W. (1985). Aging and retrieval of words in semantic memory. *J. Gerontol.* 40, 71–77. doi: 10.1093/geronj/40.1.71

Brown, C., and Hagoort, P. (1993). The processing nature of the N400: evidence from masked priming. *J. Cognit. Neurosci.* 51, 34–44. doi: 10.1162/jocn.1993.5.1.34

Burke, D. M. (1997). Language, aging, and inhibitory deficits: evaluation of a theory. *J. Gerontol. B. Psychol. Sci. Soc. Sci.* 52, P254–P264. doi: 10.1093/geronb/52B.6.P254

Chapleau, M., Wilson, M. A., Potvin, K., Harvey-Langton, A., Montembeault, M., and Brambati, S. M. (2017). Word reading aloud skills: their positive redefinition through ageing. *J. Res. Read.* 40, 297–312. doi: 10.1111/1467-9817.12065

Clahsen, H., and Reifegerste, J. (2017). "Morphological processing in old-age bilinguals," in *Bilingualism: A Framework for Understanding the Mental Lexicon,* eds M. Libben, M. Goral, and G. Libben (Amsterdam: Benjamins), 217–248.

Cohen-Shikora, E. R., and Balota, D. A. (2016). Visual word recognition across the adult lifespan. *Psychol. Aging* 31, 488–502. doi: 10.1037/pag0000100

Connor, L. T., Spiro, A., Obler, L. K., and Albert, M. L. (2004). Change in object naming ability during adulthood. *J. Gerontol. B* 59, P203–P209. doi: 10.1093/geronb/59.5.P203

Coughlin, C., Fiorentino, R., Royle, P., and Steinhauer, K. (2019). Sensitivity to inflectional morphology in a non-native language: evidence from ERPs. *Front. Comm*. 5: 888. doi: 10.3389/fcomm.2019.00021. [Epub ahead of print].

Curzietti, M., Bonnefond, A., Staub, B., Vidailhet, P., and Doignon-Camus, N. (2017). The effects of age on visual expertise for print. *Brain Lang.* 169, 48–56. doi: 10.1016/j.bandl.2017.03.001

Deacon, D., Hewitt, S., Yang, C.-M., and Nagata, M. (2000). Event-related potential indices of semantic priming using masked and unmasked words: evidence that the N400 does not reflect a post-lexical process. *Cogn. Brain Res.* 9, 137–146. doi: 10.1016/S0926-6410(99)00050-6

Derouesné, C., Poitreneau, J., Hugonot, L., Kalafat, M., Dubois, B., and Laurent, B. (1999). Le Mini-Mental State Examination (MMSE): un outil pratique pour l'évaluation de l'état cognitif des patients par le clinicien. Version française consensuelle [The Mini-mental State Examination (MMSE): a practical method for grading the cognitive state of patients for the clinician. Consensual French version]. *Presse Méd.* 28, 1141–1148. Available online at: https://docobook.com/le-mini-mental-state-examination-mm-s-e-un-outil.html

Devlin, J. T., Jamison, H. L., Matthews, P. M., and Gonnerman, L. M. (2004). Morphology and the internal structure of words. *Proc. Natl. Acad. Sci. U.S.A.* 101, 14984–14988. doi: 10.1073/pnas.0403766101

Duñabeitia, J. A., Perea, M., and Carreiras, M. (2009). Masked translation priming effects with highly proficient simultaneous bilinguals. *Exp. Psychol.* 57, 98–107. doi: 10.1027/1618-3169/a000013

Falkenstein, M., Yordanova, J., and Kolev, V. (2006). Effects of aging on slowing of motor-response generation. *Int. J. Psychophysiol.* 59, 22–29. doi: 10.1016/j.ijpsycho.2005.08.004

Feldman, L. B., and Prostko, B. (2002). Graded aspects of morphological processing: task and processing time. *Brain Lang.* 81, 12–27. doi: 10.1006/brln.2001.2503

Feyereisen, P. (1997). A meta-analytic procedure shows an age-related decline in picture naming: comments on Goulet, Ska, and Kahn (1994). *J. Speech Lang. Hear. Res.* 40, 1328–1333. doi: 10.1044/jslhr.4006.1328

Fjell, A. M., Westlye, L. T., Grydeland, H., Amlien, I., Espeseth, T., Reinvang, I., et al. (2013). Critical ages in the life course of the adult brain: nonlinear subcortical aging. *Neurobiol. Aging* 34, 2239–2247. doi: 10.1016/j.neurobiolaging.2013.04.006

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* 12, 189–198. doi: 10.1016/0022-3956(75)90026-6

Forster, K. I. (1998). The pros and cons of masked priming. *J. Psycholinguist. Res.* 2, 203–233. doi: 10.1023/A:1023202116609

Fox, J., and Weisberg, S. (2011). *An R Companion to Applied Regression*. R Package Version (Thousand Oaks, CA: Sage), 2.0–19.

Frauenfelder, U. H., and Schreuder, R. (1992). "Constraining psycholinguistic models of morphological processing and representation: The role of productivity," in *Yearbook of Morphology*, eds G. Booij and J. van Merle (Dordrecht: Kluwer), 165–183.

Gelman, A., and Su, Y.-S. (2018). *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models*. R package version 1.10-1. Available online at: https://CRAN.R-project.org/package=arm.

Giffard, B., Desgranges, B., Kerrouche, N., Piolino, P., and Eustache, F. (2003). The hyperpriming phenomenon in normal aging: a consequence of cognitive slowing? *Neuropsychology* 17, 594–601. doi: 10.1037/0894-4105.17.4.594

Good, C. D., Johnsrude, I. S., Ashburner, J., Henson, R. N., Friston, K. J., and Frackowiak, R. S. (2001). A voxel-based morphometric study of ageing in 465 normal adult human brains. *Neuroimage* 14, 21–36. doi: 10.1006/nimg.2001.0786

Gordon, J. K., and Kurczek, J. C. (2014). The aging neighborhood: phonological density in naming. *Lang. Cogn. Process.* 29, 326–344. doi: 10.1080/01690965.2013.837495

Grieder, M., Crinelli, R. M., Koenig, T., Wahlund, L.-O., Dierks, T., and Wirth, M. (2012). Electrophysiological and behavioral correlates of stable automatic semantic retrieval in aging. *Neuropsychologia* 50, 160–171. doi: 10.1016/j.neuropsychologia.2011.11.014

Hedden, T., and Gabrieli, J. D. (2004). Insights into the ageing mind: a view from cognitive neuroscience. *Nat. Rev. Neurosci.* 5, 87–96. doi: 10.1038/nrn1323

Jacob, G., Heyer, V., and Veríssimo, J. (2018). Aiming at the same target: a masked priming study directly comparing derivation and inflection in the second language. *Int. J Bilingual.* 22, 619–637. doi: 10.1177/1367006916688333

James, L. E., and Burke, D. M. (2000). Phonological priming effects on word retrieval and tip-of-the-tongue experiences in young and older adults. *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 1378–1391. doi: 10.1037//0278-7393.26.6.137

Kavé, G., and Goral, M. (2017). Do age-related word retrieval difficulties appear (or disappear) in connected speech? *Neuropsychol. Dev. Cogn. B Aging Neuropsychol. Cogn.* 24, 508–527. doi: 10.1080/13825585.2016.1226249

Kemper, S., Thompson, M., and Marquis, J. (2001). Longitudinal change in language production: effects of aging and dementia on grammatical complexity and propositional content. *Psychol. Aging* 16, 600–614. doi: 10.1037//0882-7974.16.4.600

Kiefer, M., and Brendel, D. (2006). Attentional modulation of unconscious "automatic" processes: evidence from event-related potentials in a masked priming paradigm. *J. Cogn. Neurosci.* 18, 184–198. doi: 10.1162/089892906775783688

Kliegl, R., Mayr, U., and Krampe, R. T. (1994). Time-accuracy functions for determining process and person differences: an application to cognitive aging. *Cognitive Psychol.* 26, 134–164. doi: 10.1006/cogp.1994.1005

Klinger, M. R., Burton, P. C., and Pitts, G. S. (2000). Mechanisms of unconscious priming: I. Response competition not spreading activation. *J. Exp. Psych. Learn. Mem. Cogn.* 26, 441–455. doi: 10.1037/0278-7393.26.2.441

Kouider, S., and Dehaene, S. (2007). Levels of processing during non-conscious perception: a critical review of visual masking. *Phil. Trans. R. Soc. B* 362, 857–875. doi: 10.1098/rstb.2007.2093

Kouider, S., and Dupoux, E. (2007). Partial awareness creates the "illusion" of subliminal semantic priming. *Psychol. Sci.* 15, 75–81. doi: 10.1111/j.0963-7214.2004.01502001.x

Laver, G. D., and Burk, D. M. (1993). Why do semantic priming effects increase in old age? A meta-analysis. *Psychol. Aging* 8, 34–43. doi: 10.1037/0882-7974.8.1.34

Lenth, R. (2018). *Emmeans: Estimated Marginal Means, Aka Least-Squares Means*. R package version 1.3.0. Available online at: https://CRAN.R-project.org/package=emmeans.

Lima, S. D., Hale, S., and Myerson, J. (1991). How general is general slowing? Evidence from the lexical domain. *Psychol. Aging* 6, 416–425. doi: 10.1037/0882-7974.6.3.416

Lindemer, E. R., Greve, D. N., Fischl, B. R., Augustinack, J. C., and Salat, D. H. (2017). Regional staging of white matter signal abnormalities in aging and Alzheimer's disease. *Neuroimage Clin.* 14, 156–165. doi: 10.1016/j.nicl.2017.01.022

Lüdecke, D. (2017). *Sjstats: Statistical Functions for Regression Models.* R package version 0.10.2. Available online at: https://CRAN.R-project.org/package=sjstats.

Lustig, C., and Buckner, R. L. (2004). Preserved neural correlates of priming in old age and dementia. *Neuron* 42, 865–875. doi: 10.1016/j.neuron.2004.04.002

MacKay, D. G., and James, L. E. (2004). Sequencing, speech production, and selective effects of aging on phonological and morphological speech errors. *Psychol. Aging* 19, 93–107. doi: 10.1037/0882-7974.19.1.93

Madden, D. J. (1992). Four to ten milliseconds per year: age-related slowing of visual word identification. *J. Gerontol.* 47, P59–P68. doi: 10.1093/geronj/47.2.P59

Madden, D. J., Pierce, T. W., and Allen, P. A. (1993). Age-related slowing and the time course of semantic priming in visual word identification. *Psychol. Aging* 8, 490–507. doi: 10.1037//0882-7974.8.4.490

Marquis, A., and Royle, P. (2019). "Verb acquisition in monolingual and multilingual children and adults," in *Proceedings of the GALA: Language Acquisition and Development,* eds P. Guijarro-Fuentes and C. Suárez-Gómez (Newcastle upon Tyne: Cambridge Scholars Publishing), 307–324.

McKoon, G., and Ratcliff, R. (1995). Conceptual combinations and relational contexts in free association and in priming lexical decision and naming. *Psychon. Bull. Rev.* 2, 527–533. doi: 10.3758/BF03210988

Meunier, F., and Longtin, C.-M. (2007). Morphological decomposition and semantic integration in word processing. *J. Mem. Lang.* 56, 457–471. doi: 10.1016/j.jml.2006.11.005

Milin, P., Divjak, D., and Baayen, R. H. (2017). A learning perspective on individual differences in skilled reading: exploring and exploiting orthographic and semantic discrimination cues. *J. Exp. Psychol. Learn. Mem. Cogn.* 43, 1730–1751. doi: 10.1037/xlm0000410

Montembeault, M., Joubert, S., Doyon, J., Carrier, J., Gagnon, J.-F., Monchi, O., et al. (2012). The impact of aging on gray matter structural covariance networks. *Neuroimage* 63, 754–759. doi: 10.1016/j.neuroimage.2012.06.052

Morgan-Short, K., Steinhauer, K., Sanz, C., and Ullman, M. T. (2012). Explicit and implicit second language training differentially affect the achievement of native-like brain activation patterns. *J. Cogn. Neurosci.* 24, 933–947. doi: 10.1162/jocn_a_00119

Moscoso Del Prado Martín, F. (2017). Vocabulary, grammar, sex, and aging. *Cogn. Sci.* 41, 950–975. doi: 10.1111/cogs.12367

Murray, L. L. (2000). The effects of varying attentional demands on the word retrieval skills of adults with aphasia, right hemisphere brain damage, or no brain damage. *Brain Lang.* 72, 40–72. doi: 10.1006/brln.1999.2281

Myerson, J., Adams, D. R., Hale, S., and Jenkins, L. (2003). Analysis of group differences in processing speed: brinley plots, Q-Q plots, and other conspiracies. *Psychon. Bull. Rev.* 10, 224–237. doi: 10.3758/BF03196489

Myerson, J., Ferraro, F. R., Hale, S., and Lima, S. D. (1992). General slowing in semantic priming and word recognition. *Psychol. Aging* 7, 257–270.

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., et al. (2005). The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* 53, 695–699. doi: 10.1111/j.1532-5415.2005.53221.x

Neely, J. H., Keefe, D. E., and Ross, K. L. (1989). Semantic priming in the lexical decision task: roles of prospective prime-generated expectancies and retrospective semantic matching. *J. Exp. Psychol. Learn. Mem. Cogn.* 15, 1003–1019. doi: 10.1037/0278-7393.15.6.1003

Ober, B. A., and Shenaut, G. K. (2014). Repetition priming of words and nonwords in Alzheimer's disease and normal aging. *Neuropsychology* 28, 973–983. doi: 10.1037/neu0000112

Obler, L. K., Rykhlevskaia, E., Schnyer, D., Clark-Cotton, M. R., Spiro, A., Hyun, J., et al. (2010). Bilateral brain regions associated with naming in older adults. *Brain Lang.* 113, 113–123. doi: 10.1016/j.bandl.2010.03.001

Pacton, S., and Deacon, S. H. (2008). The timing and mechanisms of children's use of morphological information in spelling: a review of evidence from English and French. *Cogn. Dev.* 23, 339–359. doi: 10.1016/j.cogdev.2007.09.004

Patterson, K., Nestor, P. J., and Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nat. Rev. Neurosci.* 8, 976–987. doi: 10.1038/nrn2277

Provost, J.-S., Brambati, S. M., Chapleau, M., and Wilson, M. A. (2016). The effect of aging on the brain network for exception word reading. *Cortex* 84, 90–100. doi: 10.1016/j.cortex.2016.09.005

Quémart, P., Casalis, S., and Colé, P. (2011). The role of form and meaning in the processing of written morphology: a priming study in French developing readers. *J. Exp. Child Psychol.* 109, 478–496. doi: 10.1016/j.jecp.2011.02.008

Rastle, K., and Davis, M. H. (2008). Morphological decomposition based on the analysis of orthography. *Lang. Cogn. Process.* 23, 942–971. doi: 10.1080/01690960802069730

Ratcliff, R., Spieler, D., and McKoon, G. (2000). Explicitly modeling the effects of aging on response time. *Psychon. B. Rev.* 7, 1–25. doi: 10.3758/BF03210723

Raveh, M., and Rueckl, J. G. (2000). Equivalent effects of inflected and derived primes: long-term morphological priming in fragment completion and lexical decision. *J. Mem. Lang.* 42, 103–119. doi: 10.1006/jmla.1999.2673

Raz, N., Gunning, F. M., Head, D., Dupuis, J. H., McQuain, J., Briggs, S. D., et al. (1997). Selective aging of the human cerebral cortex observed *in vivo*: differential vulnerability of the prefrontal gray matter. *Cereb. Cortex N. Y.* 7, 268–282. doi: 10.1093/cercor/7.3.268

Raz, N., Gunning-Dixon, F., Head, D., Rodrigue, K. M., Williamson, A., and Acker, J. D. (2004). Aging, sexual dimorphism, and hemispheric asymmetry of the cerebral cortex: replicability of regional differences in volume. *Neurobiol. Aging* 25, 377–396. doi: 10.1016/S0197-4580(03)00118-0

Reifegerste, J., and Clahsen, H. (2017). Accessing morphosyntactic information is preserved at old age, except for irregulars. *Ment. Lex.* 12, 342–372. doi: 10.1075/ml.17008.rei

Reifegerste, J., Elin, K., and Clahsen, H. (2018). Persistent differences between native speakers and bilinguals: evidence from inflectional and derivational processing in older speakers. *Bilingualism* 22, 425–440. doi: 10.1017/S1366728918000615

Robert, C., and Rico Duarte, L. (2016). Semantic richness and aging: the effect of number of features in the lexical decision task. *J. Psycholinguist. Res.* 45, 359–365. doi: 10.1007/s10936-015-9352-8

Roelofs, A., and Piai, V. (2011). Attention demands of spoken word planning: a review. *Front. Psychol.* 2:307. doi: 10.3389/fpsyg.2011.00307

Royle, P., Drury, J. E., Bourguignon, N., and Steinhauer, K. (2012). The temporal dynamics of inflected word recognition: a masked ERP priming study of French verbs. *Neuropsychologia* 50, 3542–3553. doi: 10.1016/j.neuropsychologia.2012.09.007

Rvachew, S., Royle, P., Gonnerman, L. M., Stanké, B., Marquis, A., and Herbay, A. (2017). Development of a tool to screen risk of literacy delays in French-speaking children: PHOPHLO. *Can. J. Speech-Lang. Pathol. Audiol.* 41, 321–340. Available online at: https://cjslpa.ca/files/2017_CJSLPA_Vol_41/No_03/CJSLPA_Vol_41_No_3_2017_Rvachew_et_al_321-340.pdf

Seidenberg, M. S., and Gonnerman, L. M. (2000). Explaining derivational morphology as the convergence of codes. *Trends Cogn. Sci.* 4, 353–361. doi: 10.1016/S1364-6613(00)01515-1

Shewan, C. M., and Henderson, V. L. (1988). Analysis of spontaneous language in the older normal population. *J. Commun. Disord.* 21, 139–154. doi: 10.1016/0021-9924(88)90002-0

Steinhauer, K., Royle, P., Drury, J. E., and Fromont, L. A. (2017). The priming of priming: evidence that the N400 reflects context dependent post-retrieval word integration in working memory. *Neurosci. Lett*. 651, 192–197. doi: 10.1016/j.neulet.2017.05.007

Stockall, L., and Marantz, A. (2006). A single route, full decomposition model of morphological complexity: MEG evidence. *Ment. Lex.* 1, 85–123. doi: 10.1075/ml.1.1.07sto

Sylvain-Roy, S., Lungu, O., and Belleville, S. (2015). Normal aging of the attentional control functions that underlie working memory. *J. Gerontol. B. Psychol. Sci. Soc. Sci.* 70, 698–708. doi: 10.1093/geronb/gbt166

Vallet, G. T., Simard, M., Versace, R., and Mazza, S. (2013). The perceptual nature of audiovisual interactions for semantic knowledge in young and elderly adults. *Acta Psychol.* 143, 253–260. doi: 10.1016/j.actpsy.2013.04.009

Verhaegen, C., and Poncelet, M. (2013). Changes in Naming and Semantic Abilities With Aging From 50 to 90 years. *J. Int. Neuropsychol. Soc.* 19, 119–126. doi: 10.1017/S1355617712001178

Zec, R. F., Markwell, S. J., Burkett, N. R., and Larsen, D. L. (2005). A longitudinal study of confrontation naming in the "normal" elderly. *J. Int. Neuropsychol. Soc.* 11, 716–726. doi: 10.1017/S1355617705050897

Zuccolotto, P. A., Roush, E. R., Eschman, A., and Schneider, W. (2012). *E-Prime 2.0*. Sharpsburg, PA: Psychology Software Tools, Inc.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

OPINION
published: 24 May 2019
doi: 10.3389/fpsyg.2019.01155

Idioms in the World: A Focus on Processing

Elena S. Kulkova and Martin H. Fischer†*

Department of Psychology, University of Potsdam, Potsdam, Germany

Keywords: embodiment, figurative language, metaphor, idiom, processing

OPEN ACCESS

Edited by:
Gary Libben,
Brock University, Canada

Reviewed by:
Patrick Bonin,
Université de Bourgogne, France
Mila Vulchanova,
Norwegian University of Science and Technology, Norway

****Correspondence:***
Elena S. Kulkova
kulkova@uni-potsdam.de
orcid.org/0000-0002-2145-8956

†Martin H. Fischer
orcid.org/0000-0003-3818-7272

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *31 October 2018*
Accepted: *02 May 2019*
Published: *24 May 2019*

Citation:
Kulkova ES and Fischer MH (2019) Idioms in the World: A Focus on Processing. Front. Psychol. 10:1155. doi: 10.3389/fpsyg.2019.01155

Recent experiment-based psycho- and neuro-linguistic research brought new insights into language processing mechanisms and meaning representation in the brain. More specifically, it highlighted the dynamic nature of brain connections and a constant interplay between distributed neuronal circuits during meaning processing. These developments led to a shift from an amodal view, which perceives conceptual information activation as parallel to and independent from adjoining neural activation in sensorimotor circuits (Mahon and Caramazza, 2008; Meteyard et al., 2012), to the embodied cognition view that highlights the role of sensorimotor experience in the formation of flexible, distributed conceptual representations encompassing features acquired via different perceptual modalities (Fischer and Zwaan, 2008; Barsalou, 2010). Embodied cognition, therefore, suggests that conceptual knowledge and, consequently, semantic knowledge are grounded in bodily experience and situated actions (Glenberg et al., 2008; Pulvermüller, 2013). However, currently there is a tendency toward perceiving embodied and disembodied views not as mutually exclusive distinct theories, but as bridging a gap between them. The hub and spoke model and the sensory-motor model demonstrate attempts to integrate the amodal and modality-specific views (see Mahon, 2015).

A broad range of behavioral, physiological, and neuroimaging data demonstrating co-activation of language- and action-related brain areas support this claim with regard to concrete language (Binder et al., 2005; Pulvermüller et al., 2005; Barsalou, 2008; Hauk et al., 2008). However, the data are less conclusive with regard to figurative expressions, which constitute a significant part of language. One of the reasons is that figurative language subsumes a wide variety of heterogeneous phenomena (metonymy, idioms, metaphors, proverbs, hyperbole, irony) which differ syntactically (from phrasal verbs to compounds and even sentences), as well as in their properties (familiarity, ambiguity, transparency, compositionality, salience, predictability) and essential features (although both irony and hyperbole are based on cognitive contrast, it is a contrast *of kind* for irony and a contrast *in magnitude* for hyperbole; Hsiao and Lily, 2010). This diversity and complexity of non-literal language types does not allow for clear-cut and strictly defined boundaries; it has led to a distinction of non-literal phenomena not dichotomously, but along a conventionality continuum (Cacciari and Papagno, 2012).

Secondly, the linguistic phenomena, embraced by the broad term "non-literal language" have been analyzed to different degrees of detail. Specifically, different aspects of metaphor production comprehension and use have been extensively studied (Gibbs, 2008, 2015; Schmidt and Seger, 2009; Bambini et al., 2011; Gibbs and Colston, 2012; Forgács et al., 2014; Obert et al., 2014; Lai and Desai, 2016; Briner et al., 2018; Rataj et al., 2018; Reilly et al., 2019). Furthermore, metaphors represent a powerful cognitive device guided by environmental experiences, which enabled the studies of metaphor framing influences not only on linguistic communication *per se*, but also on judgments, reasoning, intentions, and actions (Robins and Mayer, 2000; Slepian et al., 2010; Thibodeau and Boroditsky, 2011, 2013; Landau et al., 2014; Marin et al., 2014; Hauser and Schwarz, 2015; Elmore and Luna-Lucero, 2017; Thibodeau et al., 2017). Despite considerable research (Gibbs and Nayak, 1989; Cacciari and Tabossi, 1993; Mashal et al., 2008; Vulchanova et al., 2011; Cuccio et al., 2014; Häuser et al., 2016; Cacciari et al., 2018), the study of idioms still leaves open for debate the questions of defining idioms or differentiating them from other types of non-literal expressions (Cacciari, 2014). One of the main confusions is in defining idioms from metaphors, as it was debated whether idiom processing is possible without constant recourse to conceptual metaphors (Owens, 2016). However, although some idioms are indeed derived from metaphors and can still be partially motivated by conceptual mappings between domains (Gibbs, 1992), idioms as a class comprising syntactically and compositionally differing phenomena (Caillies and Butcher, 2007) are divergent from metaphors. The crucial difference is that idiomatic meaning is predominantly fixed and conventional, and it can be modified but not changed when used in various contexts. Metaphoric meaning, in turn, is flexible and intricate, can be profoundly changed by the context, and therefore always requires online construction (Cacciari, 2014; Bambini et al., 2016). Distinctive neural correlates for processing of idioms (left MTG and left IFG, involved in selection-inhibition operations) and metaphors [left precentral gyrus (BA 6), linking concrete and abstract domains and the left inferior parietal lobe (IPL), executing higher-order cognitive motor functions; Fogassi and Luppino, 2005] also argue against conflating them.

Disregarding these principal differences between the two linguistic forms results in their interchangeable use (e.g., Aziz-Zadeh et al., 2006), which, in turn, may posit serious confoundment, as comprehending these figurative devices that have different mental representations engages dissimilar cognitive mechanisms; based on both semantic and structural analysis of meaning and retrieval from semantic memory during idiom processing, and focused on the conceptual models and templates underlying metaphor meaning construction. Vulchanova et al. (2019) provide a detailed overview of the models of figurative language processing.

Recent studies on processing non-literal expressions with action-related semantics reported activation of motoric brain areas during either literal (Raposo et al., 2009), metaphoric (Desai et al., 2011), or only during metaphor but not idiom processing (Cacciari et al., 2011; Desai et al., 2013). Only limited publications present evidence of sensorimotor engagement during idiomatic meaning comprehension (Boulenger et al., 2009, 2012). Overall, the studies emphasize the role of context in meaning disambiguation and suggest that an increase in abstractness of the language stimuli leads to a decrease in the sensorimotor system's involvement (especially in case of idioms). However, these results could be interpreted not in favor of idiom disembodiment, but as a demonstration of different processing schemas that idioms and metaphors employ: the dual-reference idiomatic nature enables engagement of a hybrid processing mechanism that encompasses both compositional and holistic context-based analysis during idiom comprehension (Caillies and Butcher, 2007; Boulenger et al., 2012; Cacciari and Pesciarelli, 2013). Metaphors, which retain stronger links to the original meaning of the constituent words, rely more on online mental simulation to compute complex, flexible meanings.

Engaging different processing mechanisms may result in spatially and temporally different patterns of neurocognitive involvement (Rapp et al., 2012; Yang and Shu, 2016). For example, Cacciari et al. (2011) reported no motor engagement in idiom processing, but single-pulse TMS applied at the end of sentences to register meaning-induced MEPs could be inefficient to record idiom-induced motor activation, since idioms are processed online (mentally simulated) only until the idiom is recognized, and then a switch to the non-compositional mode (retrieval from semantic memory) occurs. Lack of motor engagement during idiom comprehension can be explained by heterogeneity of idioms: e.g., Raposo et al. (2009) used highly familiar and opaque idioms, which minimized the need for mental simulation during their processing and consequently may have reduced the level of sensorimotor cortical activation. Therefore, idiomatic meaning may be less embodied compared to metaphoric meaning, but not totally disembodied.

This evidence highlights the need for a more profound exploration of properties specific to figurative language types and subtypes of each phenomenon, which could considerably benefit present-stage figurative language research and promote a better understanding of the mechanisms the human brain employs for their acquisition, production, and processing. This will provide an integrative theoretical model that can more comprehensively and consistently outline the cognitive mechanisms and neural circuitry underlying processing of heterogeneous and multifaceted figurative language. Taken together, it will inform the development of more precise neuro-cognitive models, support AI applications and enhance understanding of language processing in general.

AUTHOR CONTRIBUTIONS

EK and MF have contributed equally to this submission.

FUNDING

This study was supported by University of Potsdam, library funds.

REFERENCES

Aziz-Zadeh, L., Wilson, S. M., Rizzolatti, G., and Iacoboni, M. (2006). Congruent embodied representations for visually presented actions and linguistic phrases describing actions. *Curr. Biol.* 16, 1818–1823. doi: 10.1016/j.cub.2006.07.060

Bambini, V., Bertini, C., Schaeken, W., Stella, A., and Di Russo, F. (2016). Disentangling metaphor from context: an ERP study. *Front. Psychol.* 7:559. doi: 10.3389/fpsyg.2016.00559

Bambini, V., Gentili, C., Ricciardi, E., Bertinetto, P. M., and Pietrini, P. (2011). Decomposing metaphor processing at the cognitive and neural level through functional magnetic resonance imaging. *Brain Res. Bull.* 86, 203–216. doi: 10.1016/j.brainresbull.2011.07.015

Barsalou, L. W. (2008). "Grounding symbolic operations in the brain's modal systems," in *Embodied Grounding Social, Cognitive, Affective, and Neuroscientific Approaches,* eds G. R. Semin and E. R. Smith (New York, NY: Cambridge University Press), 9–42. doi: 10.1017/CBO9780511805837.002

Barsalou, L. W. (2010). Grounded cognition: past, present, and future. *Top. Cogn. Sci.* 2, 716–724. doi: 10.1111/j.1756-8765.2010.01115.x

Binder, J. R., Westbury, C. F., McKiernan, K. A., Possing, E. T., and Medler, D. A. (2005). Distinct brain systems for processing concrete and abstract concepts. *J. Cogn. Neurosci.* 17, 905–917. doi: 10.1162/0898929054021102

Boulenger, V., Hauk, O., and Pulvermüller, F. (2009). Grasping ideas with the motor system: semantic somatotopy in idiom comprehension. *Cereb. Cortex* 19, 1905–1914. doi: 10.1093/cercor/bhn217

Boulenger, V., Shtyrov, Y., and Pulvermüller, F. (2012). When do you grasp the idea? MEG evidence of simultanous idiom understanding. *Neuroimage* 59, 3502–3513. doi: 10.1016/j.neuroimage.2011.11.011

Briner, S. W., Schutzenhofer, M. C., and Virtue, S. M. (2018). Hemispheric processing in conventional metaphor comprehension: the role of general knowledge. *Neuropsychologia* 114, 101–109. doi: 10.1016/j.neuropsychologia.2018.03.040

Cacciari, C. (2014). Processing multiword idiomatic strings: many words in one? *Ment. Lex.* 9, 267–293. doi: 10.1075/ml.9.2.05cac

Cacciari, C., Bolognini, N., Senna, I., Pellicciari, M. C., Miniussi, C., and Papagno, C. (2011). Literal, fictive and metaphorical motion sentences preserve the motion component of the verb: a TMS study. *Brain Lang.* 119, 149–157. doi: 10.1016/j.bandl.2011.05.004

Cacciari, C., Corradini, P., and Ferlazzo, F. (2018). Cognitive and personality components underlying spoken idiom comprehension in context. An exploratory study. *Front. Psychol.* 9:659. doi: 10.3389/fpsyg.2018.00659

Cacciari, C., and Papagno, C. (2012). "Neuropsychological and neurophysiological correlates of idiom understanding: how many hemispheres are involved," *The Handbook of the Neuropsychology of Language*, ed M. Faust (Chichester, UK: Wiley-Blackwell Publishing Ltd.), 368–385. doi: 10.1002/9781118432501.ch18

Cacciari, C., and Pesciarelli, F. (2013). Motor activation in literal and non-literal sentences: does time matter? *Front. Hum. Neurosci.* 7:202. doi: 10.3389/fnhum.2013.00202

Cacciari, C., and Tabossi, P. (1993). *Idioms. Processing, Structure and Interpretation*. Hillsdale, NJ; Hove: Lawrence Erlbaum Associates.

Caillies, S., and Butcher, K. (2007). Processing of idiomatic expressions: evidence for a new hybrid view. *Metaphor Symbol* 22, 79–108. doi: 10.1080/10926480709336754

Cuccio, V., Ambrosecchia, M., Ferri, F., Carapezza, M., Piparo, F. L., Fogassi, L., et al. (2014). How the context matters. Literal and figurative meaning in the embodied language paradigm. *PLoS ONE* 9:e115381. doi: 10.1371/journal.pone.0115381

Desai, R. H., Binder, J. R., Conant, L. L., Mano, Q. R., and Seidenberg, M. S. (2011). The neural career of sensory-motor metaphors. *J. Cogn. Neurosci.* 23, 2376–2386. doi: 10.1162/jocn.2010.21596

Desai, R. H., Conant, L. L., Binder, J. R., Park, H., and Seidenberg, M. S. (2013). A piece of the action: modulation of sensory-motor regions by action idioms and metaphors. *Neuroimage* 83, 862–869. doi: 10.1016/j.neuroimage.2013.07.044

Elmore, K. C., and Luna-Lucero, M. (2017). Light bulbs or Seeds? How metaphors for ideas influence judgments about Genius. *Soc. Psychol. Pers. Sci.* 8, 200–208. doi: 10.1177/1948550616667611

Fischer, M. H., and Zwaan, R. A. (2008). Embodied language: a review of the role of the motor system in language comprehension. *Q. J. Exp. Psychol.* 61, 825–850. doi: 10.1080/17470210701623605

Fogassi, L., and Luppino, G. (2005). Motor functions of the parietal lobe. *Curr. Opin. Neurobiol.* 15, 626–631. doi: 10.1016/j.conb.2005.10.015

Forgács, B., Lukács, Á., and Pléh, C. (2014). Lateralized processing of novel metaphors: disentangling figurativeness and novelty. *Neuropsychologia* 56, 101–109. doi: 10.1016/j.neuropsychologia.2014.01.003

Gibbs R. W. Jr. (2015). Do pragmatic signals affect conventional metaphor understanding? A failed test of deliberate metaphor theory. *J. Pragmat.* 90, 77–87. doi: 10.1016/j.pragma.2015.05.021

Gibbs R. W. Jr. (ed.) (2008). *The Cambridge Handbook of Metaphor and Thought*. New York, NY: Cambridge University Press. doi: 10.1017/CBO9780511816802

Gibbs R. W. Jr., and Colston, H. L. (2012). *Interpreting Figurative Meaning*. New York, NY: Cambridge University Press. doi: 10.1017/CBO9781139168779

Gibbs R. W. Jr., and Nayak, N. P. (1989). Psycholinguistic studies on the syntactic behavior of idioms. *Cogn. Psychol.* 21, 100–138. doi: 10.1016/0010-0285(89)90004-2

Gibbs, R. W. (1992). What do idioms really mean?. *J Mem Lang*, 31, 485–506.

Glenberg, A. M., Sato, M., Cattaneo, L., Riggio, L., Palombo, D., and Buccino, G. (2008). Processing abstract language modulates motor system activity. *Q. J. Exp. Psychol.* 61, 905–919. doi: 10.1080/17470210701625550

Hauk, O., Shtyrov, Y., and Pulvermüller, F. (2008). The time course of action and action-word comprehension in the human brain as revealed by neurophysiology. *J. Physiol. Paris* 102, 50–58. doi: 10.1016/j.jphysparis.2008.03.013

Hauser, D. J., and Schwarz, N. (2015). The war on prevention: Bellicose cancer metaphors hurt (some) prevention intentions. *Pers. Soc. Psychol. Bull.* 41, 66–77. doi: 10.1177/0146167214557006

Häuser, K. I., Titone, D. A., and Baum, S. R. (2016). The role of the ventro-lateral prefrontal cortex in idiom comprehension: an rTMS study. *Neuropsychologia* 91, 360–370. doi: 10.1016/j.neuropsychologia.2016.09.003

Hsiao, C. H., and Lily, I. (2010). Metaphor and hyperbolic expressions of emotion in Mandarin Chinese conversation. *J. Pragmat.* 42, 1380–1396. doi: 10.1016/j.pragma.2009.09.006

Lai, V. T., and Desai, R. H. (2016). The grounding of temporal metaphors. *Cortex* 76, 43–50. doi: 10.1016/j.cortex.2015.12.007

Landau, M. J., Oyserman, D., Keefer, L. A., and Smith, G. C. (2014). The college journey and academic engagement: how metaphor use enhances identity-based motivation. *J. Pers. Soc. Psychol.* 106, 679. doi: 10.1037/a0036414

Mahon, B. Z. (2015). What is embodied about cognition? *Lang. Cogn. Neurosci.* 30, 420–429. doi: 10.1080/23273798.2014.987791

Mahon, B. Z., and Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *J. Physiol. Paris* 102, 59–70. doi: 10.1016/j.jphysparis.2008.03.004

Marin, A., Reimann, M., and Castaño, R. (2014). Metaphors and creativity: direct, moderating, and mediating effects. *J. Consum. Psychol.* 24, 290–297. doi: 10.1016/j.jcps.2013.11.001

Mashal, N., Faust, M., Hendler, T., and Jung-Beeman, M. (2008). Hemispheric differences in processing the literal interpretation of idioms: converging evidence from behavioral and fMRI studies. *Cortex* 44, 848–860. doi: 10.1016/j.cortex.2007.04.004

Meteyard, L., Cuadrado, S. R., Bahrami, B., and Vigliocco, G. (2012). Coming of age: a review of embodiment and the neuroscience of semantics. *Cortex* 48, 788–804. doi: 10.1016/j.cortex.2010.11.002

Obert, A., Gierski, F., Calmus, A., Portefaix, C., Declercq, C., Pierot, L., et al. (2014). Differential bilateral involvement of the parietal gyrus during predicative metaphor processing: an auditory fMRI study. *Brain Lang.* 137, 112–119. doi: 10.1016/j.bandl.2014.08.002

Owens, J. (2016). The lexical nature of idioms. *Lang. Sci.* 57, 49–69. doi: 10.1016/j.langsci.2016.05.004

Pulvermüller, F. (2013). How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. *Trends Cogn. Sci.* 17, 458–470. doi: 10.1016/j.tics.2013.06.004

Pulvermüller, F., Hauk, O., Nikulin, V. V., and Ilmoniemi, R. J. (2005). Functional links between motor and language systems. *Eur. J. Neurosci.* 21, 793–797. doi: 10.1111/j.1460-9568.2005.03900.x

Raposo, A., Moss, H. E., Stamatakis, E. A., and Tyler, L. K. (2009). Modulation of motor and premotor cortices by actions, action words and action sentences. *Neuropsychologia*, 47, 388–396.

Rapp, A. M., Mutschler, D. E., and Erb, M. (2012). Where in the brain is nonliteral language? A coordinate-based meta-analysis of functional magnetic resonance imaging studies. *Neuroimage* 63, 600–610. doi: 10.1016/j.neuroimage.2012.06.022

Rataj, K., Przekoracka-Krawczyk, A., and van der Lubbe, R. H. (2018). On understanding creative language: the late positive complex and novel metaphor comprehension. *Brain Res.* 1678, 231–244. doi: 10.1016/j.brainres.2017.10.030

Reilly, M., Howerton, O., and Desai, R. H. (2019). Time-course of motor involvement in literal and metaphoric action sentence processing: A TMS study. *Front. Psychol.* 10:371. doi: 10.3389/fpsyg.2019.00371

Robins, S., and Mayer, R. E. (2000). The metaphor framing effect: metaphorical reasoning about text-based dilemmas. *Discourse Process.* 30, 57–86. doi: 10.1207/S15326950dp3001_03

Schmidt, G. L., and Seger, C. A. (2009). Neural correlates of metaphor processing: the roles of figurativeness, familiarity and difficulty. *Brain Cogn.* 71, 375–386. doi: 10.1016/j.bandc.2009.06.001

Slepian, M. L., Weisbuch, M., Rutchick, A. M., Newman, L. S., and Ambady, N. (2010). Shedding light on insight: priming bright ideas. *J. Exp. Soc. Psychol.* 46, 696–700. doi: 10.1016/j.jesp.2010.03.009

Thibodeau, P. H., and Boroditsky, L. (2011). Metaphors we think with: the role of metaphor in reasoning. *PLoS ONE* 6:e16782. doi: 10.1371/journal.pone.0016782

Thibodeau, P. H., and Boroditsky, L. (2013). Natural language metaphors covertly influence reasoning. *PLoS ONE* 8:e52961. doi: 10.1371/journal.pone.0052961

Thibodeau, P. H., Hendricks, R. K., and Boroditsky, L. (2017). How linguistic metaphor scaffolds reasoning. *Trends Cogn. Sci.* 21, 852–863. doi: 10.1016/j.tics.2017.07.001

Vulchanova, M., Milburn, E., Vulchanov, V., and Baggio, G. (2019). Boon or Burden? The role of compositional meaning in figurative language processing and acquisition. *J Logic Lang Information*, 1–29. doi: 10.1007/s10849-019-09282-7

Vulchanova, M., Vulchanov, V., and Stankova, M. (2011). Idiom comprehension in the first language: a developmental study. *Vigo Int. J. Appl. Linguist.* 8, 207–234.

Yang, J., and Shu, H. (2016). Involvement of the motor system in comprehension of non-literal action language: a meta-analysis study. *Brain Topogr.* 29, 94–107. doi: 10.1007/s10548-015-0427-5

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Communication

ORIGINAL RESEARCH
published: 03 June 2019
doi: 10.3389/fcomm.2019.00017

Lexical and Frequency Effects on Keystroke Timing: Challenges to a Lexical Search Account From a Type-To-Copy Task

Laurie Beth Feldman[1,2]*, *Rick Dale*[3] *and Jacolien van Rij*[4]

[1] Department of Psychology, The University at Albany, SUNY, Albany, NY, United States, [2] Haskins Laboratories, New Haven, CT, United States, [3] Department of Communication, University of California, Los Angeles, Los Angeles, CA, United States, [4] Department of Artificial Intelligence, University of Groningen, Groningen, Netherlands

We explore how timing in identical keystroke sequences that form a stem morpheme are influenced by linguistic knowledge by manipulating lexical status and morphological complexity of words in a type-to-copy task. Starting from the second keystroke, we find that average keystroke latency within a stem morpheme varies according to whole-word frequency (Experiment 1) and lexicality defined by compatibility of the upcoming suffix (e.g., IZE vs IST) with the stem (e.g., NORMAL) that forms the target string (e.g., RENORMALIZE vs. RENORMAL; RENORMALIZE vs. RENORMALIST in Experiments 2 and 3, respectively). Further, although lexical and frequency effects persist over the string as a whole, non-linear mixed-effects regressions reveal position varying lexical effects on keystroke latencies within the stem morpheme. In addition, whole word frequency effects on the first keystroke were present. These results challenge hierarchical accounts of production with modular motor programs where the same letter sequence (for a morpheme) is realized independently of and only after lexical access to the full word in which the letters occur (cf. Crump and Logan, 2010a; Logan and Crump, 2011).

Keywords: type-to-copy task, written production, morpheme, keystroke trajectory, lexicality, typing, letter position effects

Edited by:
Gonia Jarema,
Université de Montréal, Canada

Reviewed by:
Laura Barca,
Italian National Research Council (CNR), Italy
Jianfeng Yang,
Shaanxi Normal University, China

Correspondence:
Laurie Beth Feldman
lfeldman@albany.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *07 November 2018*
Accepted: *15 April 2019*
Published: *03 June 2019*

Citation:
Feldman LB, Dale R and van Rij J (2019) Lexical and Frequency Effects on Keystroke Timing: Challenges to a Lexical Search Account From a Type-To-Copy Task. Front. Commun. 4:17. doi: 10.3389/fcomm.2019.00017

INTRODUCTION

In most accounts of language processing it is assumed that access to one's knowledge about a word is stored in a mental lexicon and that lexical access to activate the requisite linguistic knowledge is necessary to execute behavior relevant to particular experimental tasks. In this framework, the mental lexicon is treated as a repository of lexical knowledge and access based on orthographic or phonological form generally is conceptualized as all or none. Nonetheless, time to access knowledge about a particular word will vary across words due to its frequency and this effect of whole word frequency gets interpreted as a reflection of the layout or organization of word representations in the lexicon. In essence, recognition of a word is conceptualized as a search through the repository whose duration depends on the manner in which it is organized and lexical retrieval is treated as all or none.

In more dynamic lexical frameworks word meaning is not stored and accessed from a form (e.g., Elman, 2004, 2009; Jones and Mewhort, 2007; Milin et al., 2017, 2018). This general approach does not conceptualize a word as an independent representation within a mental store. Instead, the knowledge that underlies productive and receptive language use reflects the typicality and distinctiveness of the meaning and form-based properties of a word with respect to context (e.g., other words). This includes both those words that are physically present and those that are not. The emphasis is more on how we learn and use language rather than on the content of localized representations for individual words (cf. Christiansen and Chater, 2016). The implication is that in a dynamic framework, the processing of a word (or morpheme or other linguistic unit) is not governed by time to execute one isolated event such as retrieving its entry in the mental lexicon because word units are not typically processed independently from one another and from other "levels" of structure (Spivey et al., 2005; Spivey and Dale, 2006; Spivey, 2008). In essence, interdependent orthographic, phonological, lexical and semantic properties that emerge over time are essential, and form the basis for a more dynamic lexical framework that involves extensive interactivity (cf. Seidenberg and McClelland, 1989).

Words are processed more quickly and more accurately than non/pseudowords, and higher frequency words are processed more quickly and more accurately than lower frequency words in a variety of experimental tasks. Interpretations of these properties nicely capture the two differing perspectives on how we use what we know about words. In the lexical repository framework, word status depends on attesting its presence in a mental lexicon and any effect of frequency reflects ease of lexical access, often conceptualized as the work of a counting mechanism that keeps track of number of prior exposures and organizes the lexical entries or the activation thresholds for particular entries according to frequency (Baus et al., 2013). Interpretations of the effect of neighborhood density—meaning number of words that differ by a single letter or phoneme from the target word—diverge from the structural interpretation for lexical status as a word and of frequency. Instead, the effect of neighborhood density reflects convergent patterns of activation and sometimes competition based on similarity with many other entries within the lexicon.

Baayen et al. (2016) have delineated how word frequency effects are, in fact, much more nuanced and not straightforwardly characterized in terms of all or none lexical access. For starters whole-word frequency values vary depending on the type of corpus on which frequencies were counted. These days some are web-based such as the Corpus of Contemporary American English (COCA, Davies, 2008), Facebook (Herdagdelen and Marelli, 2017) or Google (Brants and Franz, 2006). Others are based on subtitle frequency (Brysbaert and New, 2009). Frequency measures from different corpora tend to be correlated with each other although there are systematic differences that reflect modality (written, spoken) as well as register (formal, spontaneous). More importantly, whole word frequency measures tend to be correlated with other measures that describe letter strings. But these correlations are not restricted to real words. Letter length, orthographic neighborhood density, and more semantic measures such as emotional valence and arousal and semantic diversity and dispersion can often correlate with processing of properties of non-word sequences as well (Baayen et al., 2006). More interesting is that the correlation between various corpus-based measures of frequency and other purportedly less structural word measures tends to vary across corpora (Baayen et al., 2016). For example, an effect on processing of valence based on a subtitle corpus is stronger than from a corpus based on conversation while an effect of multiple senses or meanings is better predicted from a subtitle than in the spoken BNC corpus[1].

Most challenging for the repository account of frequency is that discrepancies with respect to frequency across corpora are not uniform across all words. Frequency estimates for high frequency words are less subject to distortions based on pockets of high usage or burstiness than are lower frequency words. In fact, some have argued that once burstiness and the concomitant contextual diversity are taken into account, the contribution of word frequency as a predictor in simple processing tasks is severely attenuated (Adelman et al., 2006, 2008).

Repository and dynamic lexical accounts of whole-word frequency invite different predictions about the role of frequency in a type-to-copy production task where target words are visually presented. Consistent with the repository account, it has been asserted that control processes in a type-to-copy production task are organized hierarchically with multiple encapsulated levels such that production constraints at one level may be impervious to constraints at another. Logan and Crump (2011) model control process for (production by) typing in terms of an outer and inner loop that are hierarchically nested. Accordingly, retrieval or selection of a particular word occurs in the outer loop while the inner loop initiates the letter and keystroke sequence for each word designated by the outer loop. Attention to an available visual template to copy is reported to be more important than visual or kinesthetic feedback in this typing task regardless of whether one types with all ten fingers or with a more limited set because the organization of keystrokes is an inner loop task and the value of a template is to the outer loop (Rieger and Bart, 2016).

Based on the usefulness of kinesthetic feedback to the inner loop and visual feedback to the outer loop in this model, Logan and Crump (2011) claim that the outer loop passes along lexical knowledge about the motor program to the inner loop but does not know about keystroke sequencing in the inner loop. In their hierarchical and sequential framework, any lexical effects on the inner loop should be constant across keystroke positions because component keystrokes are activated in parallel once a word is retrieved from the lexicon (Crump and Logan, 2010a,b; Logan and Crump, 2010). Note that after typing a word, latencies to retype a probed position have been interpreted to suggest that the activation that underlies the benefit of repetition is graded across

[1]When it comes to predicting reaction times (RTs) in a lexical decision or a naming task, Baayen et al. (2016) interpret the ostensible superiority of the subtitle corpus as reflecting a confound of frequency with other variables that affect processing in those particular time-limited tasks. By comparison, eye-tracking measures while reading prose from English novels are better predicted from the written British National Corpus (BNC) than from subtitles (Hendrix, 2015).

positions and stronger earlier in a word (Logan et al., 2016). However, reaction times to type the probe letter also were faster for first position and decreased at later positions so this finding in isolation is more difficult to interpret as consistent with parallel activation without sequential execution of constituent letters (see their Table 4 and Figure 8). Accordingly, it may be more cautious to retain the option of a systematic reduction in keystroke latencies as one progresses through the word (Rumelhart and Norman, 1982).

Manipulations of lexicality based on the legal or illegal combination of real morphemes permit one to explore where lexical and frequency effects arise in the course of producing a word in a type to copy task. Similarly, manipulations of the ratio of words to pseudowords in an experiment may affect the degree to which keystroke latencies decrease in word final positions. Novel in our typing study is that differences in latency between finger movements to particular keys are controlled for by comparing words and pseudowords that share a morpheme and thus a letter sequence. As a result, the influence on keystroke latency of the distance a finger must move to the key (Fitts, 1954) as well as the decision of which fingers to use (Hick, 1952; Hyman, 1953) are weakened if not fully eliminated. An account of production that entails retrieving from the lexical repository a typing motor program and control for its execution in a loop that is immune to lexical influences would need to assert that the relationship between keystrokes (e.g., [N]ORMAL) should be stable across words that contain that letter sequence or morpheme. In many analyses the first letter [N] is not included because the initial keystroke of a word tends to be disproportionally longer than the others. We follow that practice here and indicate it by the bracket notation [N].

Essential in this framework is that strings may differ with respect to time for lexical retrieval but, because keystroke execution in the inner loop occurs automatically, the relative timing for the same sequence of letters such as [N]ORMAL within strings such as RENORMALIZE vs. *RENORMALIST should not vary with lexicality (Shaffer, 1975; Gentner et al., 1988). One outcome that is more compatible with a dynamic than with a repository framework is that predictability in various types of linguistic contexts interacts with keystroke dynamics (plausibly in complex ways) so as to influence the manner in which a word is produced or recognized. In essence the dynamic account, but not the retrieval account, would not only be compatible with but would anticipate non-linear changes in position by keystroke latencies with manipulations of lexicality or whole-word frequency.

Online typing tasks with dependent measures based on the execution of keystrokes within a morpheme have been useful to track the interdependence of morpheme and word structure in production (Gagné and Spalding, 2014; Feldman et al., 2017). In the present study, we attenuate the role of retrieving or selecting the target word by presenting the target word visually and then ask whether whole word frequency and other lexical effects are restricted to the outer loop or whether inner loop measures associated with execution of constituent letters are sensitive to linguistic factors as well.

We track three basic measures of typing performance. All are sensitive to which finger moves and to what key. The simplest and best investigated is latency to key contact for the initial keystroke in a word (K_1). This measure is assumed by Logan and Crump (2011) to reflect response preparation and initiation and it is known to be sensitive to whole-word frequency (West and Sabban, 1982; Inhoff, 1991; Pinet et al., 2016) and also to word length (Gagné and Spalding, 2014). Second is average latency between keystrokes for letters within a letter string or inter keystroke interval (IKSI). This measure is interpreted to reflect execution of the motor plan and is sensitive to bigram frequency (Pinet et al., 2016) and again to word length (Gagné and Spalding, 2014, 2016). Finally, it is possible to examine average keystroke latency by position within a stem morpheme or word. As a rule when typing text, the timing between keystrokes (IKSI) is faster at the end of a word and faster for a higher frequency letter or letter sequence (e.g., bigram, trigram) than for a less frequent letter sequence (Gagné and Spalding, 2016; Pinet et al., 2016).

Overall, IKSIs covary with multiple measures of predictability defined within as well as across words (van Rij et al., 2019a). For example, even the word "the," perhaps the most typed word in all of English, can be sped up or slowed down slightly depending on how predictable it is in context (van Rij et al., 2019a). The implication is that one should account for differences in keystroke latencies by letter and bigram frequency before attempting to examine their interaction with various word properties including frequency or lexicality. At issue is whether position by keystroke latencies for identical constituent letters vary systematically according to lexical properties (letter length, orthographic neighborhood density, whole word frequency) of the words or stems in which they appear. The study we report utilizes effects on processing of a stem morpheme that varies systematically according to the other morphemes with which it appears. What we find does not support a characterization of morphological processing that assumes decomposition, or a characterization that emphasizes processing of a stem in isolation from the other constituents with which it typically appears. Effects of lexicality that arise from an incompatible affix positioned after the stem such as RENORMALIST are potentially informative in this regard.

We report the results of three experiments that use non-linear mixed-effects regressions to compare measures based on variation of keystroke timing. Comparisons focus on the same stem morpheme (e.g., NORMAL) in a variety of morphological contexts in an online typing-to-copy task. The key comparison in Experiment 1 is stem keystroke latencies between words that differ in whole word frequency such as NORMALLY and NORMALCY. In Experiment 2 the critical comparison is between strings that differ in affixation and resultant lexical status such as *RENORMAL which is not a word and RENORMALIZE which is and in Experiment 3 it is prefixed and suffixed strings that differ in lexical status such as *RENORMALIST and RENORMALIZE. Regardless of any preoccupation with morphological decomposition, morphological knowledge in production is particularly interesting to examine in its own right as it can provide a framework that highlights interactions of lexical effects on typing speed across the keystroke positions within a letter string.

In fact, the term lexical status is deceptive because letter strings can vary in their degree of wordiness. As consideration of *RENORMAL and *RENORMALIST demonstrate, a letter string need not be in full compliance or in full violation with lexical knowledge about word formation. Patterning can be graded. Readers know what these particular combinations of morphemes would mean, even though neither is an attested word in the language. Because we compare conditions where stems repeat, comparisons of the same pattern of keystrokes in more or less predictable morphological contexts allow us to answer various questions about the time course over which linguistic knowledge emerges in a production task. For example, differences between keystroke measures in RENORMALIZE vs. RENORMALIST speak to when an upcoming lexical deviation becomes evident on IKSI. Differences between keystroke measures in NORMALIZE vs. RENORMALIZE, words that differ in whole word frequency because of a prefix show how long the effect of a prefix and the concomitant reduction in whole word frequency persist on IKSI.

In summary, our focus on morphemes in production provides a method to control for differences in keystroke latencies by letter and bigram and permits an examination of the interaction of lexicality and various word properties like whole-word frequency and word length on morphological processing. Our focus on keystroke latency measures shows that these measures vary across types of morphological combinations in a type-to-copy production task. The results do not have strong compatibility with an account based on search and all-or-none access to a lexical repository. For example, work based on time to initiate first keystroke or average keystroke latency (less the initial keystroke) are fully compatible with models that assume lexical access and retrieval of a motor program before initiation of a motor response. In contrast, keystroke-to-keystroke latencies that vary across positions of a letter sequence within a word raise the possibility of a more dynamic option as when the timing variation in the execution of keystroke movements for a stem vary systematically with the lexical or morphological properties of the string as a whole.

These keystroke sequence latencies across positions within a word pose a challenge to the notion that response preparation based on lexical access and retrieval of a motor program is completed in its entirety before the initiation of keystroke movements. Similar claims for ongoing (re) assessment of lexicality have been made in the domain of comprehension when the lexical determination for a morphologically simple letter string is indicated by the velocity profile of mouse movement (Barca and Pezzulo, 2012, 2015).

Of particular relevance in the type-to-copy task is whether variation in IKSIs as one progresses through a morpheme or a word reflect a systematic and continuous updating based on lexical status, predictability, whole word frequency and perhaps other linguistic factors or whether effects remain constant over letter positions because decisions about which keystrokes to activate occur before movement to keystrokes begins. In support of prolonged linguistic influences on typing measures, Gagné and Spalding (2014, 2016) have reported a slowing in IKSI at the boundary between morphemes in a word and this effect is sensitive to the semantic consistency of the critical morpheme to the meaning of the full word in which it appears (Libben et al., 2012, 2014; Gagné and Spalding, 2016). Semantic influences on keystroke differences within a word can vary by position. They generally appear at but may appear earlier than the stem boundary. For the time being, we ask whether differences between conditions that vary according to the combination of morphemes are salient when aligned to the beginning of a morpheme stem and leave analyses aligned to the end for future work. Stem-initial alignment invites a focus on anticipatory influences whereas reductions with stem final alignment could reflect a later wrapping up that maximizes the semantics of the stem morpheme with respect to the suffixes with which it can combine. Both could be semantic in nature but to differing degrees and ultimately are worthy of consideration. Most relevant for the time being is evidence that lexical influences can be revised and updated during the course of producing a word.

METHODS

To investigate lexical and frequency effects on keystroke measures, we conducted three on-line typing experiments. All used the same procedure with slightly different materials all of which consisted of triplets formed around a morpheme stem.

Participants

Recruitment and Payment

For each experiment, we aimed to recruit 100 (target $N = 100$) participants on Amazon's Mechanical Turk (AMT; requester.mturk.com). We report below procedures for discarding any participants and data, so that our N in each experiment was lower than this target. Participants were given the opportunity to participate once in the task for a payment of $1.25.

Participant Demographics

We restricted recruitment to the US, to people with at least 100 prior approved tasks, with an approval rate of 95%. One hundred participants were recruited for each of the three experiments. A small number of payments on AMT were rejected, no more than 2 per dataset. For example, participants who submitted a response on AMT had their work rejected due to lack of an appropriate payment confirmation number. The modal age range was 26–34 (∼50% of participants). For highest education level, the two most common responses included 48 with "high-school" and 43 with "bachelors." All participants reported gender: 62% of participants reported gender as male, 38% female. One hundred percent reported QWERTY keyboards. Eighty one percent were right-handed, and 1 participant reported ambidextrous.

Participant Time and Exclusion

The mode of the time required to complete the task was approximately 5 min. The average was 15 min. Data from participants that appeared not to be complete were excluded.

We also excluded participants who did not appear to have typed responses to all trials in the experiment. This left data for 258 participants (85 in Experiment 1, 87 in Experiment 2, and 86 in Experiment 3), each of whom typed 25–58 words correctly.

Procedure

We adapted an Internet typing task that facilitates rapid data collection through AMT (Vinson et al., under revision). This online task uses a JavaScript framework to track IKSIs while participants type a word that is displayed on the screen. The framework records milliseconds associated with each keystroke, and tracks which key was pressed. This offers an easy-to-use type-to-copy task that rapidly crowdsources large amounts of typing data. An additional resource of the interface and raw data, in addition to our supplementary data, can be found online[2].

After ∼2 s (ISI), participants were presented a word and a textbox. Their cursor was automatically focused on the textbox so they did not have to use the mouse. They then typed words back and hit ENTER to move on. At any point an error was made, the interface reported it to the participants and began the next trial. About 7 practice items preceded the experimental items. The interface is depicted in **Figure 1A**. Informed consent was incorporated into the instructions for the task. See **Figure 1B**.

Materials

The final set of experimental materials for each experiment included about 60 words. The materials for Experiment 1 consisted of only 57 triplets formed around a shared stem morpheme because three stem triples contained spelling errors and were eliminated from all analyses. The triplet based on the stem "NORMAL" provides an example throughout this report. In Experiment 1, for each triplet, one item was the stem such as NORMAL. The second and third items were a legal stem-suffix combination of that same stem such as NORMAL-LY and NORMAL-CY, with the former occurring with a higher frequency than the latter.

In Experiment 2, for each triplet, one item was a legal stem-suffix combination such as NORMAL-IZE. The second was a legal prefixed version of that same stem plus suffix sequence such as RE-NORMAL-IZE. The third was a nonword formed by combining the stem morpheme of the former with an incompatible prefix such as RE-NORMAL. The *RENORMAL constraint required that we present different materials than in Experiment 1.

In Experiment 3, for each triplet, one item was a legal stem-suffix combination such as NORMAL-IZE. The second was a legal prefixed version of that same stem plus suffix sequence such as RE-NORMAL-IZE. The third was a nonword formed by combining the prefix and the stem morpheme of the former with an incompatible suffix such as RE-NORMAL-IST. Suffixes in the prefixed word and prefixed nonword condition were matched triple by triple for length but not number of syllables with a median of 4 and a range of 2 and 4. Prefixes in those two conditions had a median of 2 and ranged only between 2 and 3. Median suffix length was 4 letters but varied between 2 and 4.

In no case did the stem morpheme undergo a spelling change when affixed as in the derivation of SEVERITY from SEVERE where the final E in SEVERE gets dropped before affixation. Each participant viewed and typed one formation from the stem morpheme. Members of each stem triplet were distributed across three different lists and presented to different participants. Each participant viewed and typed a total 57 unique words. For example across lists, the same stem morpheme NORMAL appeared in different morphological contexts e.g., NORMALIZE, RENORMALIZE, and RENORMALIST.

Table 1 presents the conditions (types of word structures) that were tested in each experiment. Materials are listed in **Appendix A**.

RESULTS

To consolidate presentation of the data and facilitate comparisons of a measure across experiments and types of word structures, we present the results of the three experiments in parallel. Materials for Experiments 2 and 3 only differed in whether or not nonwords were suffixed. Thus, the experimental stimuli in Experiments 2 and 3 were almost identical in that they were derived from the same stems. Experiment 2 included three additional stems, however. Experiment 1 only shared four stems with Experiment 2 and 3.We focus on keystroke latencies across positions within a stem that is nested within a letter string because systematic variation poses a challenge to the notion that response preparation based on lexical access and retrieval of a motor program is completed in its entirety before the initiation of keystroke movements.

Preprocessing of Data

The trials were terminated at the end of the presented string or when a typing error was made. Incorrect trials were not included in the analysis (2,298 trials out of 14,712; 15.6%). The overall accuracy varied slightly between the three experiments (86.9, 84.8, and 81.6% respectively, for Experiments 1, 2, and 3) but this variation is likely attributable to variations in word length (average number of characters: 8.3, 8.9, and 9.9 respectively, for Experiments 1, 2, and 3): As a rule, longer words were typed less accurately than shorter words. After excluding all incorrect trials, 70 trials (out of 12,414; 0.56%) were excluded because the first keystroke latency was longer than 6,000 ms, and another 7 trials were excluded because one or more later keystroke latencies were longer than 3,000 ms (0.056%).

Analysis

A range of measures can be used to investigate when lexical and orthographic information would become available during the time course of typing a word. These include the first keystroke latencies, the sum of keystroke latencies for letters in the stem (normalized for length), all keystroke latencies for letters in the stem, and the trajectory meaning the keystroke latencies by letter position within the string. In this paper we will present only the three most important measures, namely

[2]https://github.com/racdale/keystroke-timing-feldman-dale-van-rij

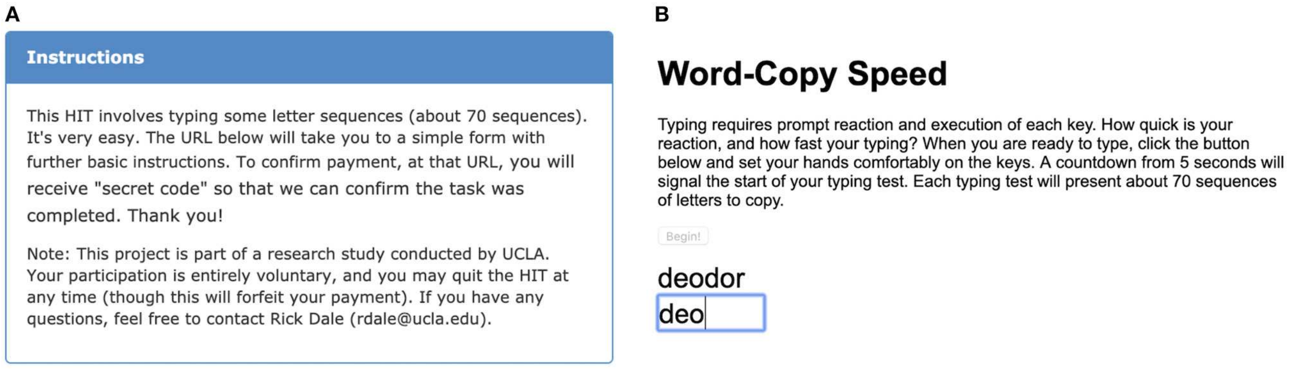


FIGURE 1 | (A) Screen interface on Amazon Mechanical Turk. **(B)** Informed consent in interface instructions.

TABLE 1 | Overview of the experimental conditions in the three experiments, with example stimuli derived from the stem "normal."

Experiment 1	Experiment 2	Experiment 3
Stem "normal"	Suffix "normalize"	Suffix "normalize"
Suffix-HF "normally"	Prefix-Suffix "renormalize"	Prefix-Suffix "renormalize"
Suffix-LF "normalcy"	Nonword (Prefix) "renormal"	Nonword (Prefix-Suffix) "renormalist"

the first *keystroke latency (K_1)*, which is purported to reflect lexical access (Crump and Logan, 2010b), *the keystroke latencies for the stem as a whole less* K_1, which reflects execution of the motor program to type the stem and is most compatible with the morphological word recognition literature where stem processing is the primary focus, and the *trajectory of keystroke latencies by letter position*, which has the potential to provide more insight into keystroke by keystroke execution of the motor program to type the stem. To reiterate the logic, systematic differences in the same (series of) keystrokes depending on co-occurring morphemes within a word call into question the claim that all keystrokes are activated in parallel without regard to lexical context.

The typing measures were analyzed with Generalized Additive Mixed Modeling (GAMM; Wood, 2017), a mixed-effects regression approach that allows a non-linear relation between the measure and the covariates (see for introductions Wieling, 2018; van Rij et al., 2019a,b). The data were analyzed in R version 3.4.4 (2018-03-15; (R Core Team, 2018)) using the package *mgcv* version 1.8-24 (Wood, 2017) for modeling GAMMs and the package *itsadug* version 2.3 (van Rij et al., 2017) for evaluation and visualization of the results. The data of the three experiments were separately analyzed with similar statistical models, unless stated otherwise. We used an iterative backward-fitting model comparison procedure for determining the best-fitting model, but we also inspected the summary statistics and visualizations of the effects to verify the conclusions (cf. Wieling, 2018; van Rij et al., 2019b). The models were fitted using the maximum likelihood optimization score.

To investigate the effect of lexical frequency, we used the frequencies in the Google Books Corpus (Total word counts for English, version 20120701; 543,081 words and 6,640,052,764 tokens). We excluded the occurrences in books published before 1950 (leaving 541,040 words and 3,345,974,073 tokens). The frequency was converted to frequency per million words, and log-transformed to approach a normal distribution. We additionally calculated a measure of orthographic similarity, OLD50, which is the average Levenshtein distance between a word from the experiment and its 50 nearest neighbors in the Google Books Corpus of books after 1950 (cf. Yarkoni et al., 2008), using the R package *vwr* version 0.3.0 (Keuleers, 2013). The OLD50 scores were log-transformed.

First Keystroke Latencies

Figure 2 shows average log frequency per condition in each experiment. **Figure 3** presents the grand averages of the first keystroke latencies (K_1) for the three experiments and the conditions within each experiment. On average, the first keystroke latency is 850 ms. These are the most noticeable differences: in Experiment 1, the low frequency suffix words ("Suffix-LF") seem to start with a longer first keystroke latency than the other two conditions (855 ms vs. 812 ms Stem/816 ms Suffix-HF); in Experiments 2 and 3 the Prefix-Suffix words (882 and 858 ms, respectively) seem to start with a longer first keystroke latencies than the Suffix words (856 and 830 ms, respectively). Faster K_1 latencies for higher as compared to lower frequency words replicate reports in the typing literature (Crump and Logan, 2010a,b; Logan and Crump, 2011).

Differences Between Conditions

For the analyses, the keystroke latencies were transformed with an inverse transformation to approach normality (−1,000/keystroke latency). GAMM analyses were performed on the data for each experiment separately, with random intercepts included for the stem-triplets (i.e., words in three different conditions that were derived from the same stem), and for

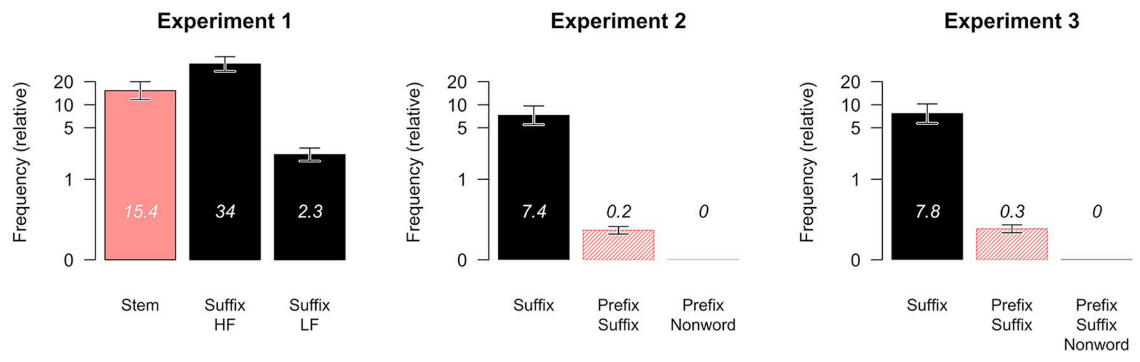


FIGURE 2 | Mean (±1 SE) of the relative frequencies (frequency per million; on log scale) of the stimuli in Experiments 1, 2, and 3.

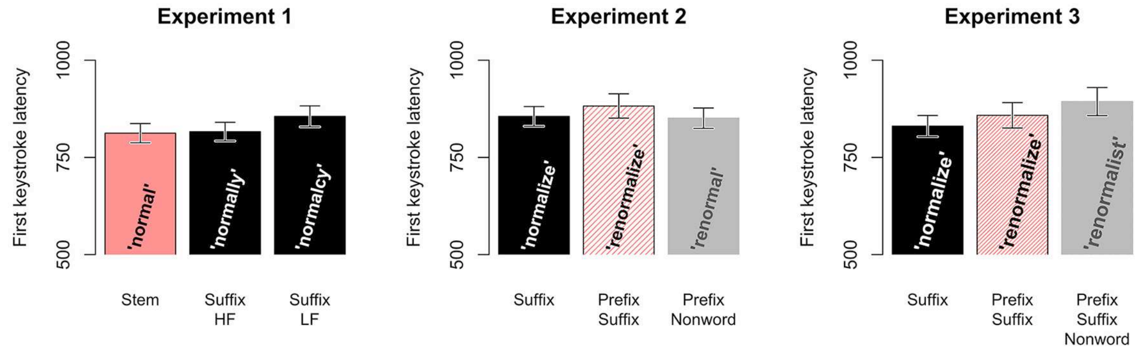


FIGURE 3 | Mean (±1 SE) of the first keystroke latencies in Experiments 1, 2, and 3.

the first keystroke letter, and with by-participant non-linear random smooth's for *Trial* (i.e., the position of the word in the presentation sequence) to capture fluctuations in typing latencies over the course of the experiment that could cause autocorrelation in the residuals (Baayen and Milin, 2010). Model comparisons were utilized to determine whether the predictor *Condition*, which marks the three different experimental conditions, improved the model.

The GAMM analyses indicated that the first keystroke latencies in Experiment 1 were significantly influenced by condition [$\chi^2_{(2)} = 12.32$; $p < 0.001$; $\Delta AIC = 24.4$]. Latencies on Stem words were shorter than in high-frequency suffix (Suffix-HF) words ($\beta_{Stem} = -0.0191$, SE $= 0.0083$; t-value $= -2.30$; $p = 0.021$). More importantly, the first keystroke latencies in low-frequency suffix words (Suffix-LF) were significantly longer than in high-frequency suffix words ($\beta_{S-LF} = 0.0224$, SE $=$ 0.0084; t-value $= 2.65$; $p = 0.008$). The first keystroke latencies in Experiment 2 were also significantly influenced by condition [$\chi^2_{(2)} = 3.05$; $p = 0.048$; $\Delta AIC = 5.18$]: latencies for Prefix-Suffix words were longer than for Suffix words ($\beta_{P-S} = 0.040$, SE $= 0.020$; t-value $= 2.03$; $p = 0.042$). While, the first keystroke latencies in *Experiment 3* did not differ between Prefix-Suffix words and Suffix words ($\beta_{P-S} = 0.013$, SE $= 0.023$; t-value $= 0.57$; $p > 0.1$), the effect of Condition [$\chi^2_{(2)} = 9.58$; $p <$ 0.001; $\Delta AIC = 18.11$] was reliable for the difference between length matched words like RENORMALIZE and nonwords like RENORMALIST.

The difference between high- and low-frequency suffixed words formed from the same stem in Experiment 1 is consistent with reports that word frequency influences the first keystroke latency such that first keystroke takes less time in the higher frequency suffix words than in the lower frequency suffix words. Words in those conditions differed in their frequency, but not in word length. The difference between the high-frequency suffix words and the Stem words, on the other hand, suggests that word length may play a role. The first keystrokes of the Stem words take less time to produce than in high-frequency suffixed words, but their frequency is lower on average than the high-frequency suffixed words (see **Figure 2**). Thus, both frequency and word length can affect first keystroke latencies. Based on these conclusions, we would expect to find a difference in the first keystroke latencies in Experiments 2 and 3 between the Suffix words and the Prefix-Suffix words, because they differ in word length and in frequency (Prefix-Suffix words are longer and have a lower average frequency than Suffix words, see **Figure 2**). However, this difference reached significance in Experiment 2, but not in Experiment 3.

Another factor that may influence the first keystroke latencies is the morphological complexity of the words: Words composed only of a stem should be easier to process than those affixed words with a suffix or prefix, and those affixed words with both (Prefix-Suffix) should be most difficult. To isolate an effect of frequency from morphological complexity, we investigated the effects of frequency on suffixed words by combining materials across experiments.

Effect of Frequency

To investigate the effects of whole word frequency on the first keystroke latencies more directly, we combined all Suffix-words from the three Experiments into one analysis (i.e., the conditions "Suffix-LF" and "Suffix-HF" from Experiment 1, and the conditions called "Suffix" from Experiments 2 and 3 –the black bars in **Figure 1**). Random intercepts for participants, stem-triplets, and the typed letters were included in the GAMM model, along with a by-participant random smooth for log Frequency and a by-participant random slope for OLD50. We included the predictor *Experiment* to test for differences between the experiments, and non-linear smooth's for whole word *Frequency*, and *OLD50*, our measure of orthographic distance from the 50 most similar words. Word length was not included as a predictor, because the experimental stimuli did not exhibit sufficient variation in stem length (range 5–6).

The statistical model indicated that the first keystroke latencies for suffixed words in Experiment 1 were significantly faster than for those in Experiments 2 and 3 ($\beta_{1-2} = -0.080$, SE = 0.043, *t*-value = 1.88, $p = 0.061$; $\beta_{1-3} = 0.095$, SE = 0.035, $t = 2.71$, $p = 0.007$). This could reflect at least in part the inclusion of pseudowords in the latter two experiments. In addition, the effect of Frequency was significantly different from zero [$F_{(1.00,5121.559)} = 11.74$; $p < 0.001$], and linear (edf = 1.00); an edf (effective degrees of freedom of the smooth term) of 1 indicates a straight line)[3]. Visualization of the effect of frequency across experiments indicated that Suffix-words with lower frequency result in a longer first keystroke latency than Suffix-words with higher frequency. This is illustrated in **Figure 4** (Left panel). OLD50 did not contribute to the model (see **Figure 4**, Right panel). Finally, the interaction between Experiment and Frequency did not improve the model [a model without the interaction resulted is a lower ML score (ΔML = 10.04), fewer degrees of freedom (Δdf = 4), and a lower AIC (ΔAIC = 18.11)].

In summary, we have replicated the effect of whole word frequency on initial keystroke latency (in this experiment only marginally significant) and extended it to words composed of a stem and a suffix. Thus, we add a finding from a production task to the literature showing a robust effect of whole word frequency thereby complementing those identified in recognition tasks. In the next section we explore keystroke latencies to the stem independent from any effect of initial keystroke. We ask whether lexical effects are evident in processes associated with a purportedly encapsulated inner loop that controls keystroke execution.

[3] An edf (effective degrees of freedom of the smooth term) of 1 indicates a straight line.

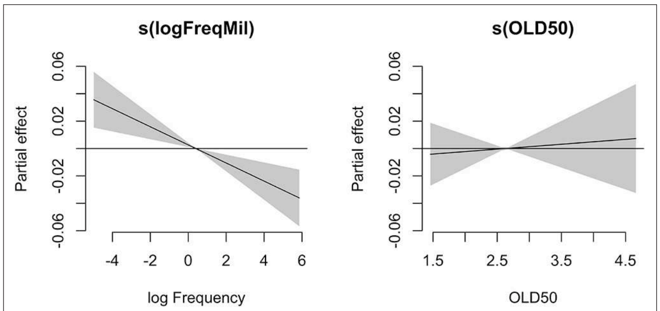


FIGURE 4 | Partial effect estimates from the GAMM model. **Left:** Effect of frequency on the first keystroke latencies of Suffix-words. **Right:** (Non-significant) effect of OLD50 (orthographic similarity measure) on the first keystroke latencies of Suffix-words.

Keystroke Latencies on the Stem (Normalized for Length)

The keystrokes following the initial keystroke, were typed considerably faster than the first. The average latency of the later keystrokes was 194 ms, which is considerably shorter than the first keystroke latency of 850 ms. With deference to the visual word recognition literature for morphology, we analyzed the keystroke latencies of the stem. To avoid redundancy with the analysis of the first keystroke of the word, in the stem latencies we always excluded the first keystroke of the stem. Further, the stem latency was normalized for stem length by dividing the sum of the latencies by the number of keystrokes (i.e., $stemRT = \frac{5}{n}\sum_{i=2}^{n} k_i$, with n the stem length and k_i the keystroke latencies). **Figure 4** presents the stem latencies for the conditions of the three experiments. Similarly to the first keystroke latencies (**Figure 1**), Experiment 1 shows the longest latencies for the low-frequency suffix words, and the shortest latencies for the stem words, but note that the difference between the stem and the high-frequency suffix words seems to be larger here than in the first keystroke latencies. In Experiments 2 and 3 the difference between the Prefix-Suffix and Suffix words also is more systematic than in the first keystroke latencies.

We adhered to the same procedure in analyzing the stem latencies as with the first keystroke latencies: first we investigated the differences between the conditions in each experiment, and then we combined the Suffix words from all experiments to investigate the effect of frequency and orthographic neighborhood density.

Differences Between Average Stem Keystroke Latencies Across Conditions

For the analyses, the keystroke latencies were again transformed with an inverse transformation to approach normality (−1,000/keystroke latency). GAMM analyses were performed on the data for each experiment separately, with random intercepts included for the word-triplets (i.e., words in three different conditions that were derived from the same stem), and with by-participant non-linear random smooth's for *Trial* (i.e., the position of the word in the course of the experiment) to capture fluctuations in typing latencies that can result in autocorrelation in the residuals. Model comparisons were conducted to evaluate whether inclusion of *Condition* improved the model.

The GAMM analyses indicated that the average stem keystroke latencies in *Experiment 1* were significantly influenced by condition [$\chi^2_{(2)} = 101.96$; $p < 0.001$; $\Delta AIC = 206.96$]: latencies in Stem words were shorter than in high-frequency suffix (Suffix-HF) words ($\beta_{Stem} = -0.0787$, SE = 0.0099; *t*-value = −7.98; $p < 0.001$), but the average stem keystroke latencies in low-frequency suffix words (Suffix-LF) was significantly longer than in high-frequency suffix words ($\beta_{S-LF} = 0.0641$, SE = 0.0101; *t*-value = 6.37; $p < 0.001$). The average stem keystroke latencies in *Experiment 2* were also significantly influenced by condition [$\chi^2_{(2)} = 98.08$; $p < 0.001$; $\Delta AIC = 198.45$]: latencies in Prefix-Suffix words were longer than Suffix words ($\beta_{P-S} = 0.077$, SE = 0.011; *t*-value = 6.90; $p < 0.001$), but latencies in Prefix-Nonwords were significantly shorter than Suffix words ($\beta_{PN-S} = -0.080$, SE = 0.011; *t*- value = −7.27; $p < 0.001$). Similarly, for *Experiment 3* the average stem keystroke latencies were significantly influenced by condition [$\chi^2_{(2)} = 179.90$; $p < 0.001$; $\Delta AIC = 355.62$): latencies in Prefix-Suffix words were longer than Suffix words ($\beta_{S-PS} = -0.107$, SE = 0.011; *t*-value = −9.39; $p < 0.001$), but shorter than the Prefix-Suffix nonwords ($\beta_{PSN-PS} = 0.111$, SE=0.012; *t*- value = 9.53; $p < 0.001$).

Different from the first keystroke latencies, the average stem latencies show reliable and systematic differences among all conditions. These differences are much stronger than the differences found with the first keystroke measure and indicate lexical effects on keystroke dynamics independent of the first keystroke. To investigate the effect of frequency and orthographic neighborhood density, again we analyzed the Suffix words.

Effect of Frequency

To ascertain the effects of frequency on the stem latencies, we combined all Suffix-words from the three Experiments in one analysis (i.e., the conditions "Suffix-LF" and "Suffix-HF" from Experiment 1, and the conditions called "Suffix" from Experiments 2 and 3—the black bars in **Figure 5**). As above, random intercepts for participants, and stem-triplets were included in the GAMM model, and a by-participant random smooth for log Frequency along with a by-participant random slope for OLD50. We included the predictor *Experiment* to test for differences between the experiments, and non-linear smooth's for *Frequency*, and *OLD50*, a measure of orthographic distance from the 50 most similar words that is similar in function to measures of bigram and trigram frequency. Word length was not included as a predictor, because the experimental stimuli did not show sufficient variation in word length.

The statistical model indicated that the stem latencies did not differ between the experiments even though the ratio of words to pseudowords varied. However, the effects of Frequency [edf = 1.77; $F_{(1.770,5016.967)} = 12.34$; $p < 0.001$] and OLD50 [edf = 1.00; $F_{(1.000,5016.967)} = 7.18$; $p < 0.01$] were significantly different from zero and followed a linear trend. As there were some words included which were not found in the corpus, we ran the model again without those extremely low frequencies to verify whether the effects of Frequency and OLD50 could be attributed to those outliers. In this new model the effects of Frequency [$F_{(1.000,4496.278)} = 13.71$; $p < 0.001$] and OLD50 [$F_{(1.000,4496.278)} = 6.28$; $p = 0.012$] remained significantly different from zero. Visualization of the effects depicted that for the same keystrokes, Suffix-words with lower frequency resulted in a longer stem latency than Suffix-words with higher frequency (**Figure 6**, Left panel). In addition, Suffix words with a shorter average distance (i.e., higher orthographic similarity) with the 50 most similar

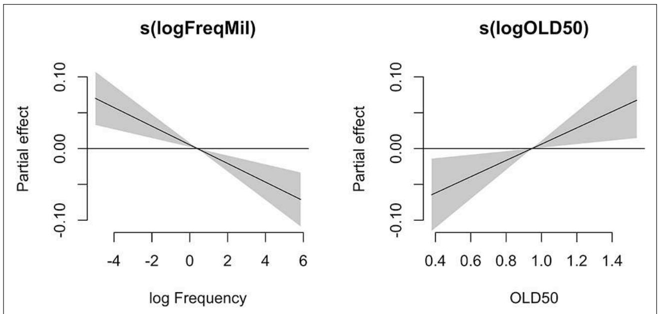


FIGURE 6 | Partial effect estimates from the GAMM model. **Left:** Effect of frequency on the first keystroke latencies of Suffix-words. **Right:** Effect of OLD50 (orthographic similarity measure) on the first keystroke latencies of Suffix-words.

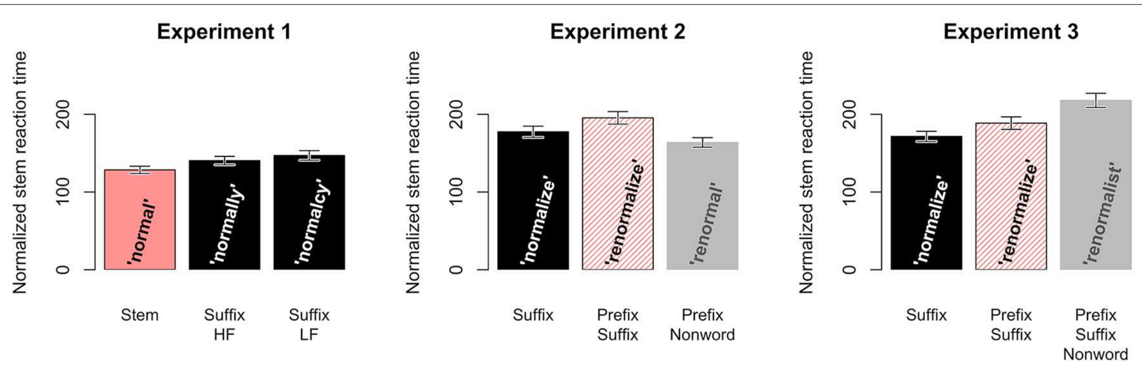


FIGURE 5 | Mean (±1 SE) of the normalized stem latencies in Experiments 1, 2, and 3.

words resulted in shorter stem latencies than words with less orthographically similar neighbors (**Figure 6**, Right panel).

These combined analyses indicate that when typing the stem, both lexical knowledge based on whole word frequency as well as orthographic knowledge based on similarity with other words is available and influences the typing speed. Whereas, effects of orthographic knowledge on keystroke latencies have been documented frequently in the past, effects of whole word frequency on keystroke latencies after the initial keystroke have not.

Keystroke Latencies by Position in Stem

If variation in non-initial keystroke latencies across positions within a stem depends on the letter string within which it is nested then preparation and retrieval of a motor program cannot be completed in its entirety before the initiation of keystroke movements. Here we use GAMMs (Wood, 2017) to compare the trajectories for keystroke timing across stem position in lexical and morphological contexts formed around the same stem. We expect decreases in latencies across position and ask whether rate of keystroke execution decreases uniformly across stems that differ with respect to position of affix(es) and the lexical status of the particular combination when position within the word is held constant. Here, we examine the time course of morphological effects in production when keystrokes are aligned to stem onset but abutting morphemes differ.

Perhaps most obvious in **Figure 7** is that keystroke latencies are not uniform across the stem and further, they vary according to the structure of the string in which the stem appears. Consistent with previous reports of slowing around the boundary between morphemes, increased latencies are visible at the onset of the stem after a prefix in Experiments 2 and 3. Stem-suffix boundary effects are difficult to detect, however, at least in part because of variation in suffix length. As a rule, stem by position latencies decrease both in the absence of a suffix and, to a lesser degree, in its presence. More interestingly, whole word frequency contributions introduced by manipulations of suffix are evident not only when typing the letters of the suffix but also in the course of typing the letters of the stem (Experiment 1). Evidently, production of keystroke latencies for the stem are not independent of the context in which it appears. For example in Experiment 1, the possibility of competing suffixes such as CY and LY as one transitions out of a stem such as NORMAL seems to offset the typical speeding up that occurs as one approaches the end of the word (e.g., positions 8–12). In both Experiments 2 and 3, stem latencies are faster in the production of a suffixed only word than in the production of that same string when accompanied by a prefix. See **Figure B1** (**Appendix B**) for the same non-initial keystroke latencies aligned on the *offset* of the stem.

Results such as these highlight some of the ways in which processing of the stem is interdependent with that of the affixes with which it co-occurs. We examine this interaction in more detail below because it may identify a potential weakness of an account of morphological processing restricted to the stem, and an account of typing where keystrokes are executed in series irrespective of emerging lexical or non-lexical context based on the particular combination of morphemes which accompany the stem.

Analyses of Keystroke Trajectories Aligned to Stem Onset

In order to further examine whether rate of keystroke execution was stable or decreased uniformly across positions in the word, we compared keystroke by position within stems that appeared in contexts composed of various combinations of affixes in *keystroke trajectory* analyses. The keystroke latencies again were transformed with a log-transformation to approach normality and we then excluded the first keystroke of the word from the analyses. We included the following predictors in our statistical models: *Condition*, which marks the three different morphological structures within an experiment, *Keystroke Position*, which captures the position of the keystroke within the stem relative to its onset, *Key*, indicating the particular letter that was typed, and *Stimulus*, describing the word-triplets, i.e., words in three different conditions that were derived from the same stem, and *Participant*. GAMM analyses were performed on the data for each experiment separately, with non-linear random smooth's included for Keystroke Position by Stimulus, and non-linear random smooth's for Keystroke Position by Participant by Condition, and a random intercept for Key, capturing the variation in typing caused by the different letters. The models were fitted using the smoothing parameter estimation method fREML (fast restricted maximum likelihood) for estimating the smoothing parameters, because the data were too large to use ML (maximum likelihood). As a consequence, the model comparisons may be less reliable. Therefore, we used both the model summary information and a model-comparison procedure to determine whether the predictors Condition and Keystroke Position and their interaction explained significantly more variance in the data than the baseline model with only random effects included.

For *Experiment 1*, we ran the GAMM analysis on the keystroke latencies in the stem (excluding the first keystroke), because the Stem words did not contain a suffix. **Figure 8** (left panel) illustrates the estimated effects from the best-fitting statistical model. The model with the interaction between Condition and Keystroke Position included had a lower AIC value than the model without the interaction (ΔAIC $= 8.69$), but the fREML scores were not significantly different. Inspection of the estimated effects suggests that when typing the stem, there was no significant difference in keystroke latencies by position between the conditions Suffix-HF and Suffix-LF, although the latencies were faster when typing the Stem-words than in the other two suffixed conditions. Any distinctiveness of Suffix-HF keystrokes arose mainly at the end of the stem (see **Figure 8**, left). In addition, the summary statistics indicate a non-linear trend for typing keystrokes in the Stem words, that was significantly different from zero [$F_{\text{Stem}(3.732,20421.281)} = 4.28$; $p < 0.01$], but nothing comparable for Suffix words. This outcome indicates that there is no difference between timing of keystroke positions for the stems in HF and LF productions as one produces the word.

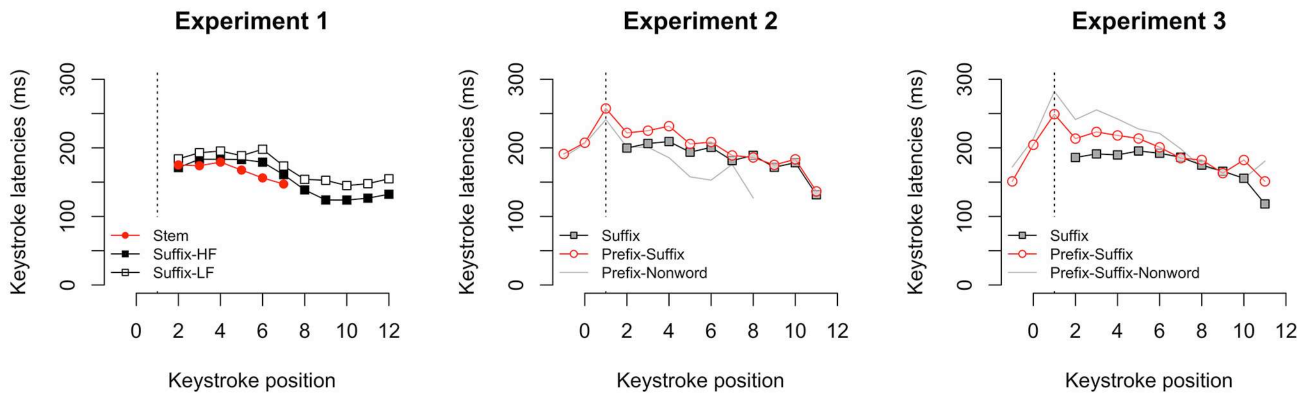


FIGURE 7 | Mean keystroke trajectories (excluding the first keystroke) aligned to the onset of the stem for all conditions in Experiments 1, 2, and 3. The onset of the stem is indicated with a dashed vertical line.

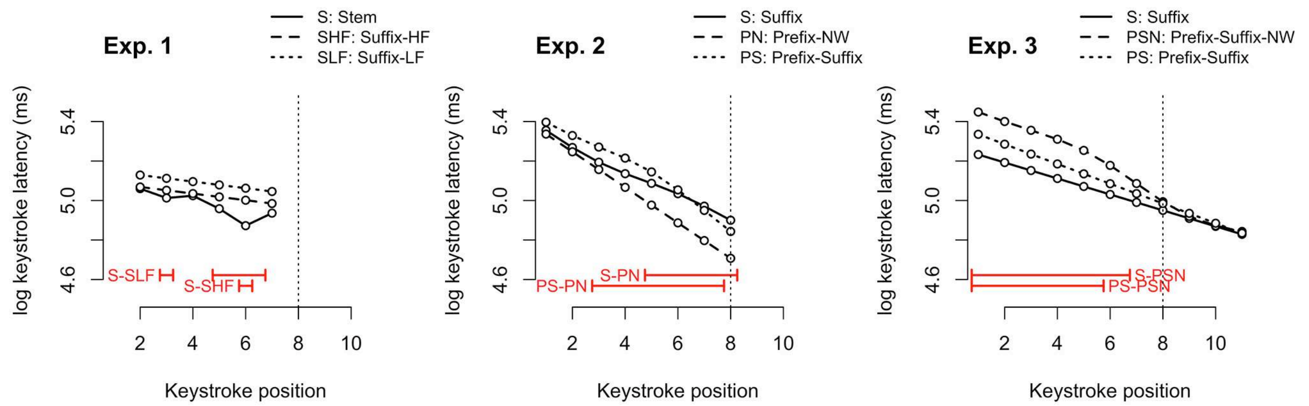


FIGURE 8 | Estimates from the GAMM models fitting the keystroke trajectories (excluding the first keystroke) aligned to the onset of the stem for all conditions ([stem, stem + suffix low frequency; stem + suffix high frequency]; [stem + suffix; prefix + stem (NW); prefix + stem + suffix]; [stem + suffix; prefix + stem + suffix (NW); prefix+ stem + suffix] in Experiments 1, 2, and 3 respectively. The onset of the stem is labeled as 1. Positions with significant latency differences are indicated in red. The vertical dashed line indicates the median onset of the suffixes, which are only included in the analysis of Experiment 3.

These results fail to be consistent with the analysis of the average stem latencies where there was a significant difference between the two Suffix-conditions. It seems to be the case that the latencies on the low frequency Suffix-words are slightly longer than on the high frequency Suffix-words. The difference was not sufficient to establish significantly different keystroke trajectories but did indicate a difference when we summed the latencies across positions to calculate the average stem latencies. Note that if the program to produce all keystrokes were retrieved in parallel and executed according to the same sequencing constraints then the pattern of keystroke latencies should not differ depending on the upcoming morphemes. Either trajectories should not vary by position within the word or perhaps they should decrease uniformly in later positions within the word but an effect on the stem of an upcoming affix and the lexical acceptability of the stem-affix combination are not anticipated in a repository account.

For *Experiment 2*, we ran the GAMM analysis on the keystroke latencies in the stem only (excluding the first keystroke of the word), because the Prefix-Nonwords did not contain a suffix, and the Suffix words did not contain a prefix. Here again, model comparisons suggested that the model with the interaction between Keystroke Position and Condition explained significantly more variance than a model without this interaction [$\chi^2_{(4)} = 8.17, p = 0.003; \Delta\text{AIC} = 7.05$]. In contrast with the results of Experiment 1, the summary statistics of this model indicate that the trends over Keystroke Position for all three conditions are significantly different from zero [$F_{\text{Suffix}(2.247, 22396.844)} = 5.05$, $p < 0.01$; $F_{\text{Prefix-Suffix}(2.545, 22396.844)} = 7.94$, $p < 0.001$; $F_{\text{Prefix-Nonword}(1.019, 22396.844)} = 32.00$, $p < 0.001$]. Further inspection of the estimated effects suggests that there was no significant difference in keystroke latencies between the word conditions Suffix and Prefix-Suffix when typing the stem, but that the latencies were *faster* when typing the Prefix nonwords than in the other two word conditions (see **Figure 8**, center). An effect of lexicality on prefixed strings emerged early in the stem, which was followed by a suffix in the word conditions, but not in the nonword condition. The difference between the words and nonwords seems to point to the absence of a word final speed up as arose in the presence of competing suffixes in Experiment 1. In Experiment 3 we included nonwords with a suffix, to further probe lexicality effects.

In contrast to the analyses for Experiments 1 and 2, we included the keystrokes on the stem and the suffix when analyzing the data of *Experiment 3* (continuing to exclude the first keystroke of the word). Once again, model comparisons suggest that the model with the interaction between Keystroke Position and Condition explained significantly more variance than did a model without this interaction [$\chi^2_{(4)} = 6.19$, $p = 0.015$; $\Delta\text{AIC} = 13.02$]. The summary statistics reveal that both word conditions, Prefix-Suffix and Suffix words, were better fitted with a linear regression line (edf = 1) rather than a non-linear trend. All three conditions show a significant decrease in latencies with increasing Keystroke Position [$F_{\text{Suffix}(1.001,31387.140)} = 22.00$, $p < 0.001$; $F_{\text{Prefix-Suffix}(1.000,31387.140)} = 38.11$, $p < 0.001$; $F_{\text{Prefix-Suffix-Nonword}(3.394,31387.140)} = 21.09$, $p < 0.001$]. Inspection of the estimated effects suggests that there is no significant difference in keystroke latencies between the conditions Suffix and Prefix-Suffix when typing the stem. Importantly however, the latencies were *slower* when typing the Prefix-Suffix nonwords than in the other two conditions (see **Figure 8**, right). This finding supports the idea that nonwords are more difficult to type than words with the same stem and same morphological (affixation) structure. Note that the differences in keystroke latencies between the words and nonwords disappears at the end of the stem (see also **Figure 8**, right panel). Lexicality of the prefix-stem-suffix sequence influences early keystroke latencies but by the time that participants are typing the suffix, any effect of lexicality has dissipated.

Taken together, the analyses of all three experiments show a speed-up in latencies at the end of the word and an early lexicality effect, with slower latencies for nonwords than words on the stem, but not on the suffix.

Keystroke Variation as a Function of Whole Word Frequency

Our final insights into the production of morphologically complex words derive from analyses of the conditions under which effects of orthographic similarity interact with frequency. We again combine the two Suffix-conditions from Experiments 2 and 3 into one analysis. In contrast with the earlier analysis of the first keystroke latency and the stem latencies, we did not include the Suffix conditions from Experiment 1 in this analysis, because these showed larger variation in suffix as well as stem length, which necessarily makes aligning the trajectories more complex: Suffix words in Experiment 1 varied in the length of suffix between 1 (e.g., "jealousy") and 6 characters (e.g., "satisfactory"). Materials in Experiments 2 and 3 underwent less variation in suffix length (2–4 characters).

Random non-linear smooth's for Keystroke Position by Stimulus, Keystroke by Participant, and Frequency by Participants were included in the GAMM model, along with a random intercept for Key. We included the predictor Experiment to test for differences between the experiments, and non-linear smooth's for Keystroke Position (aligned on the onset of the stem), Frequency, and OLD50, a measure of orthographic distance with the 50 most similar words. Word length was not included as predictor, because the experimental stimuli were constructed so as to restrict variation in word length. Of most interest were potential non-linear interactions between Keystroke Position and Frequency and between Keystroke Position and OLD50. Here again, the models were fitted using fREML for estimating the smoothing parameters, but because the data were too large for using ML, the model comparisons may be less reliable. Therefore, we report the summary statistics when these provide information on the contribution of a predictor or interaction. We excluded 10 words that did not occur in the Google Books Corpus as outliers from the analyses presented below, but we verified that this did not change the results by rerunning the models on all data. As the differences are small, we present here the data without the outliers.

The interaction between Keystroke Position on the stem plus suffix and OLD50 was significantly different from zero [$F_{(3.478,16833.605)} = 4.08$; $p < 0.01$], but the interaction between Keystroke Position and Frequency made only a weak contribution to the model [$F_{(2.524,16833.605)} = 2.78$; p=0.034] (it did not reach significance with all data included). Further, there was a linear main effect for Frequency [$F_{(1.000,16833.605)} = 15.27$; $p < 0.001$] and for Keystroke Position [$F_{(3.248,16833.605)} = 5.44$; $p < 0.001$] with no effect of experiment. **Figure 9** presents the model estimates of the *partial effects* (i.e., individual model terms) for Keystroke Position, Frequency, and the interaction between Keystroke Position and Frequency on the top row. The effect of Keystroke Position showed a general decrease in keystroke latencies, along with an even steeper decrease after the stem (the stem length ranges between 4 and 8 characters, with a median stem length of 6 characters).

In the analysis of the stem trajectories of Experiment 3 suffixes are included because all conditions had suffixes. In this case, no boundary pattern was detectable in the analysis. There was an overall frequency effect, with lower word frequencies resulting in generally longer keystroke latencies. The partial interaction effect indicates that the word frequency effect is stronger when typing the affix than when typing the stem. The center panel (surrounded by a box) shows the *summed effects* for Keystroke Position and Frequency, including the partial effects in the top row and the intercept. Of particular relevance is that higher frequency words show a much steeper decrease in keystroke latencies at the end of the word than do lower frequency words. Finally, the bottom left panel shows the estimated interaction (partial effect) of Keystroke Position by OLD50. This interaction suggests that the effect of orthographic uniqueness (lower OLD50 values) increases the latencies at the end of the stem and on the suffix. For the longer words, this effect seems to be reversed around the last characters. Basically, any effect of orthographic similarity comes in later than the effects of whole word frequency.

At a minimum, it is evident that effects of frequency on keystrokes latencies persist throughout an entire word and that they are not uniform across position; as a rule, they decrease across positions in the word. Perhaps most important is that the rate of speeding up varies with whole word frequency. Stated

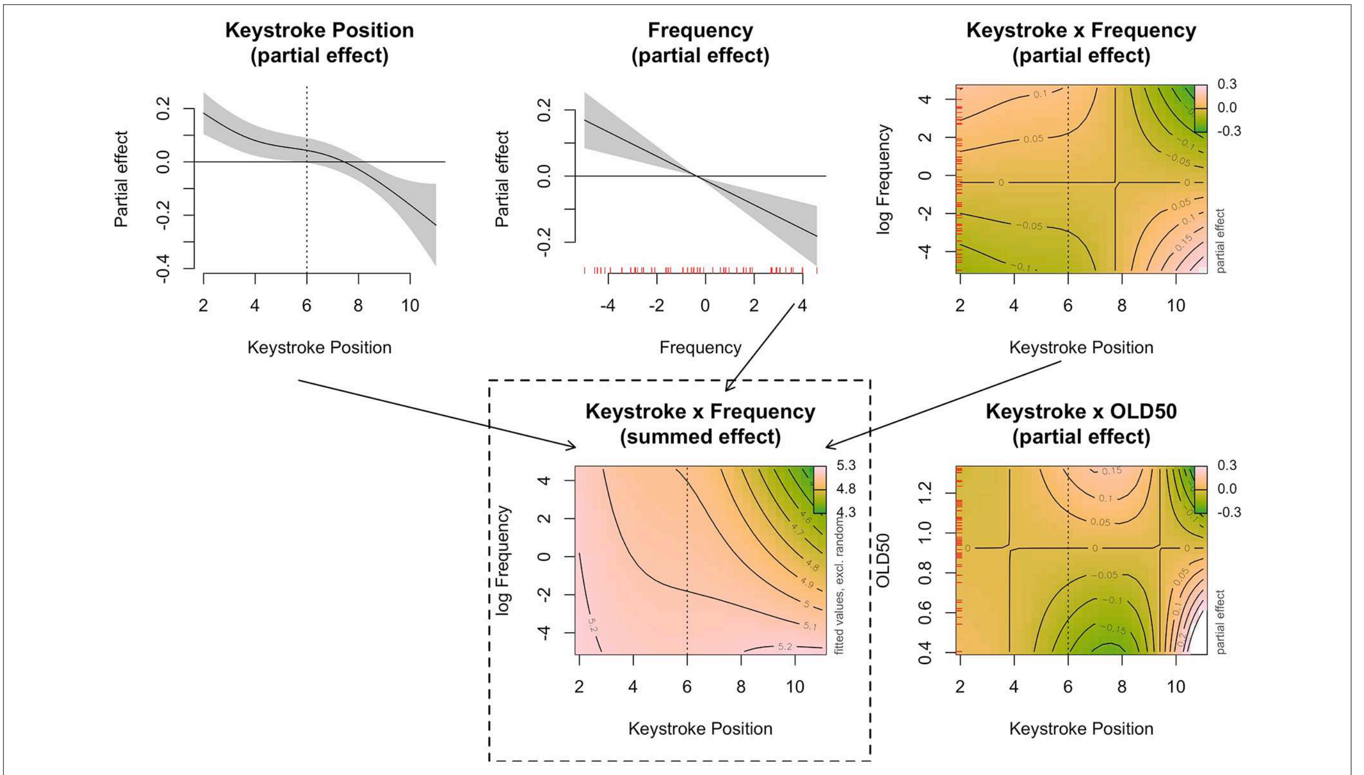


FIGURE 9 | Estimates from the best-fitting statistical models of the keystroke trajectories (excluding the first keystrokes of the words) aligned to the onset of the stem for the Suffix words in Experiments 2, and 3. The top row shows the partial effects of Keystroke Position **(left)**, Frequency **(center)**, and Keystroke Position by Frequency **(right)**. The bottom row shows the partial effect of Keystroke Position by OLD50 **(right)** and the summed effects of Keystroke Position by Frequency, with random effects excluded **(center)**.

succinctly, keystrokes latencies in higher frequency words show a more dramatic reduction at the end of the word than in lower frequency words.

DISCUSSION

In the present study we asked whether average keystroke latencies and related measures for identical constituent letters in a stem vary systematically according to their lexical properties (letter length, orthographic neighborhood density) or to those of the words in which they appear (whole word frequency). As a rule, higher frequency words were typed more quickly and more accurately than lower frequency words, and words were typed more quickly and more accurately than nonwords when length was matched. In the framework where the mental lexicon is treated as a repository of lexical knowledge and access to it is conceptualized as all or none, time to access knowledge about a particular word will vary across words due to frequency and this effect of whole word frequency gets reflected in the layout or organization of word representations in the lexicon. In essence, recognition of a word or the motor program to type it is described as a search through a repository of words where search duration depends on the manner in which the repository is organized.

A prominent recent model of expert typing is compatible with this tradition and posits two independent loops (Crump and Logan, 2010a,b; Logan and Crump, 2010, 2011). In this framework, the outer loop passes along lexical knowledge about the requisite motor program to the inner loop but is blind to keystroke sequencing or timing which are the responsibility of the inner loop. It is successful in accounting for a number of interesting effects (see Logan, 2018 for review). As we have highlighted above, the model does not predict that lexical effects that persist into the inner loop should vary across keystroke positions because component keystrokes are activated in parallel and executed in series once a word is retrieved (Crump and Logan, 2010a,b; Logan and Crump, 2010, 2011). As noted above, however, effects of retyping a probed position seem more consistent with graded activation across positions, because effects are stronger earlier in a word (Logan et al., 2016). In this case, the higher-level word unit may be activating all of the keystrokes in parallel but there is some indication that execution varies with position within the word. Similarly, degree of disruption to typical keystroke position vary according to the position of the target letter within the word (Yamaguchi and Logan, 2014).

In the present study, we provide novel evidence that activation as measured by keystroke latency does vary with position within the word and that it is not uniform across contexts when length is controlled. Rather, measures based on keystroke latencies can be influenced by stem position within the string, by string lexicality and affixation, and by similarity of the target string to other words. It is important to note that in prior discussion, Logan and Crump (2011, p. 7) do acknowledge the potential cross-talk between these loops, but argue that these relationships are unlikely to contribute substantially to explaining variation in production. Such a comparison of effect size between purported outer and inner loops is outside the scope of the present study. However, the methods we described in our three experiments may permit new investigation of these distinctions through new tasks and, importantly, new statistical models. We elaborate below.

Most novel in our study was the analysis of *keystroke trajectories* which revealed not only that rate of keystroke execution decreased across positions in the word but also that those changes were not uniform over different morphological structures and word frequencies. In order to test these effects, we took a multilevel model-building approach, integrated various item- and subject-level factors contributing to the sequence of interkey intervals, and thereby controlled for a variety of factors. Across three experiments, these analyses helped quantify subtle aspects of word production in typing.

Of particular note were the anticipatory effects of an upcoming affix on keystroke trajectories according to the lexical acceptability of the combination. In the tradition of morphological decomposition in the recognition literature, one might have expected the contribution of the stem to predominate over that of any affixes that were produced at the same time. In fact, the effect of Keystroke Position showed a general decrease in keystroke latencies, along with an even steeper decrease after the stem. Effects of whole word frequency on keystroke timing could be documented with several keystrokes measured on the stem but, here again, the trajectory analysis in **Figure 9** indicated that the word frequency effect was more pronounced when typing the affix than when typing the stem. Finally, higher frequency words showed a steeper decrease in keystroke latencies at the end of the word than did lower frequency words. Similarly, we observed an interaction (partial effect) of Keystroke Position with OLD50 such that orthographic uniqueness (lower OLD50 values) increased keystroke latencies at the end of the stem and into the suffix.

In these results, keystroke latencies are not retrieved and executed in a uniform manner. Finally, the dynamic but not the retrieval account anticipates interactions of typing measures with orthographic similarity of the target to other words or to predictability of the affix given the particular stem. Processes at different levels (visual form recognition, morphological segmentation, semantic processing, etc.) seem to be fluidly interacting throughout performance. This interaction among processes could be the mechanistic underpinning of language processing and production. Finding new echoes of this parallelism in behavioral metrics offers a promising new direction to test such predictions about styles of interaction among levels of control.

SUMMARY AND CONCLUSION

In the word recognition literature that focuses on morphological processing, the repository account typically asserts that access to the lexicon entails decomposing a morphologically complex word into its constituent morphemes by a process that is blind with respect to the semantics of the stem. Recent reports demonstrating the salience of whole word as contrasted with stem frequency in the course of morphological processing have substantially weakened this account (Baayen et al., 2007; Milin et al., 2017; Schmidke et al., 2017). Once one accesses the lexicon, one can retrieve the motor program to produce a word as by typing it, but we have demonstrated here that that process is not executed independently from its lexical and morphological properties.

We reported the results of three experiments where the critical comparison focuses on a repeated stem morpheme in a variety of morphological and lexical contexts. As in the decomposition account in word recognition, if morphological decomposition and stem access dominated processing in an online typing-to-copy task, then structures that accompanied the stem could have been ignored or played only a secondary role. In this framework one might have expected latencies for the keystrokes that comprise the stem should have been more stable over morphological and lexical contexts. On the contrary, as depicted in **Figures 5–7**, this was not the case. Rather, patterns of keystroke latencies for letters in the stem highlight the interactions of lexical and morphological effects on stem production. Further, an effect of lexicality based on an incompatible stem- suffix combination emerged while executing the keystrokes of the stem in anticipation of the upcoming deviation. The theoretical upshot is an unencapsulated parallelism—from motor control to morphological semantics, these patterns and systematicities are whispering to each other in a manner that is measurable in performance.

ETHICS STATEMENT

Data were collected online and we had no face to face contact with participants. The UCLA Institutional Review Board (UCLA IRB) has determined that the above referenced study meets the criteria for an exemption from IRB review. UCLA's Federal wide Assurance (FWA) with Department of Health and Human Services is FWA00004642.

AUTHOR CONTRIBUTIONS

RD designed the web interface for data collection. LF designed the materials. JvR created the analyses. All authors were engaged in writing the manuscript.

FUNDING

The development of this publication was supported by the Words in the World Partnership Project funded by the Social Sciences and Humanities Research Council of Canada (895-2016-1008).

REFERENCES

Adelman, J. S., Brown, G. D. A., and Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychol. Sci.* 17, 814–823. doi: 10.1111/j.1467-9280.2006.01787.x

Adelman, J. S., Brown, G. D. A., and Quesada, J. F. (2008). Modeling lexical decision: the form of frequency and diversity effects. P*sychol Rev*. 115, 214–229. doi: 10.1037/0033-295X.115.1.214

Baayen, R. H., Feldman, L. F., and Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *J. Mem. Lang.* 53, 496–512 doi: 10.1016/j.jml.2006.03.008

Baayen, R. H., and Milin, P. (2010). Analyzing reaction times. *Int. J. Psychol. Res.* 3, 12–28.

Baayen, R. H., Milin, P., and Ramscar, M. (2016). Frequency in lexical processing, *Aphasiology* 30, 1174–1220. doi: 10.1080/02687038.2016.1147767

Baayen, R. H., Wurm, H. L., and Aycock, J. (2007). Lexical dynamics for low-frequency complex words. *A* regression study across tasks and modalities. *The Mental Lexicon* 2.3, 419–463. doi: 10.1075/ml.2.3.06baa

Barca, L., and Pezzulo, G. (2012). Unfolding visual lexical decision in time. *PLoS ONE* 7:e35932. doi: 10.1371/journal.pone.0035932

Barca, L., and Pezzulo, G. (2015). Tracking second thoughts: continuous and discrete revision processes during visual lexical decision. *PLoS ONE*, 10:e0116193. doi: 10.1371/journal.pone.0116193

Baus, C., Strijkers, K., and Costa, A. (2013). When does word frequency influence written production? *Front Psychol.* 4:963. doi: 10.3389/fpsyg.2013.00963

Brants, T., and Franz, A. (2006). Web 1T 5-gram, Version 1. Philadelphia: Linguistic Data Consortium, University of Pennsylvania

Brysbaert, M., and New, B. (2009). Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behav. Res. Methods*. 41, 977–990. doi: 10.3758/BRM.41.4.977

Christiansen, M. H., and Chater, N. (2016). The now-or-never bottleneck: a fundamental constraint on language. *Behav. Brain Sci.* 39, 1–72. doi: 10.1017/S0140525X1500031X

Crump, M. J. C., and Logan, G. D. (2010a). Hierarchical control and skilled typing: Evidence for word-level control over the execution of individual keystrokes. *J. Exp. Psychol. Learn. Memory Cogn.* 36, 1369–1380. doi: 10.1037/a0020696

Crump, M. J. C., and Logan, G. D. (2010b). Warning: This keyboard will deconstruct—The role of the keyboard in skilled typewriting. *Psychon. Bull. Rev.* 17, 394–399. doi: 10.3758/PBR.17.3.394

Davies, M. (2008). *The Corpus of Contemporary American English (COCA): 410+ Million Words, 1990-Present*. Available online at: www.americancorpus.org

Elman, J. L. (2004). An alternative view of the mental lexicon. *Trends Cogn. Sci.* 8, 301–306. doi: 10.1016/j.tics.2004.05.003

Elman, J. L. (2009). On the meaning of words and dinosaur bones: lexical knowledge without a lexicon. *Cogn. Sci.* 33, 547–582. doi: 10.1111/j.1551-6709.2009.01023.x

Feldman, L. B., Vinson, D., and Dale, R. (2017). "Production of morphologically complex words as revealed by a typing task: Interkeystroke measures challenge the viability of morphemes as production units," in *Paper Presented at The Mental Lexicon*, *October* (Ottawa, ON).

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *J. Exp. Psychol.* 47, 381–391. doi: 10.1037/h0055392

Gagné, C. L., and Spalding, T. L. (2014). Typing time as an index of morphological and semantic effects during English compound processing. *Lingue E Linguaggio* 13, 241–262. doi: 10.1418/78409

Gagné, C. L., and Spalding, T. L. (2016). Effects of morphology and semantic transparency on typing latencies in english compound and pseudocompound words. *J Exp Psychol Learn Mem Cogn.* 42, 1489–95. doi: 10.1037/xlm0000258

Gentner, D. R., Larochelle, S., and Grudin, J. (1988). Lexical, sublexical, and peripheral effects in skilled typewriting. *Cogn. Psychol.* 20, 524–548. doi: 10.1016/0010-0285(88)90015-1

Hendrix, P. (2015). *Experimental Explorations of a Discrimination Learning Approach to Language Processing*. Doctoral Dissertation, University of Tuebingen.

Herdagdelen, A., and Marelli, M. (2017). Social Media and language processing: how facebookand twitter provide the best frequency estimates for studying word recognition. *Cogn. Sci.* 41, 976–995. doi: 10.1111/cogs.12392

Hick, W. E. (1952). On the rate of gain of information. *Q J Exp Psychol.* 4, 11–26. doi: 10.1080/17470215208416600

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *J Exp Psychol.* 45, 188–96. doi: 10.1037/h0056940

Inhoff, A. W. (1991). Word frequency during copytyping. *J Exp Psychol Hum Percep Perform.* 17, 478–487. doi: 10.1037/0096-1523.17.2.478

Jones, M. N., and Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychol. Rev.* 114, 1–37. doi: 10.1037/0033-295X.114.1.1

Keuleers, E. (2013). *vwr: Useful Functions for Visual Word Recognition Research*. R package version 0.3.0. Available online at: https://CRAN.R-project.org/package=vwr

Libben, G., Curtiss, K., and Weber, S. (2014). Psychocentricity and participant profiles: implications for lexical processing among multilinguals. *Front. Psychol.* 5:557. doi: 10.3389/fpsyg.2014.00557

Libben, G., Weber, S., and Miwa, K. (2012). A technique for the study of perception, production, and participant properties. *Mental Lexicon* 7, 237–248. doi: 10.1075/ml.7.2.05lib

Logan, G. D. (2018). Automatic control: How experts act without thinking. *Psychol. Rev.* 125, 453–485. doi: 10.1037/rev0000100

Logan, G. D., and Crump, M. J. C. (2010). Cognitive illusions of authorship reveal hierarchical error detection in skilled typists. *Science* 330, 683–686. doi: 10.1126/science.1190483

Logan, G. D., and Crump, M. J. C. (2011). "Hierarchical control of cognitive processes: the case for skilled typewriting," in *Psychology of Learning and Motivation*, Vol. 54, ed B. H. Ross (Burlington: Academic Press), 1–27.

Logan, G. D., Ulrich, J. E., and Lindsey, D. R. B. (2016). Different (key)strokes for different folks: how standard and nonstandard typists balance Fitts' law and Hick's law. *J Exp Psychol.* 42, 2084–2102.

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., and Baayen, R. H. (2017). Discrimination in lexical decision. *PLoS ONE* 12:e0171935. doi: 10.1371/journal.pone.0171935

Milin, P., Smolka, E., and Feldman, L. B. (2018). "Models of lexical access and morphological processing," in *The Handbook of Psycholinguistics*, eds H. Cairns and E. Fernandez (London: Whiley).

Pinet, S., and. Ziegler, J. C., and Alario, F.-X. (2016). Typing is writing: Linguistic properties modulate typing execution. *Psych Bull Rev.* 23, 1898–906. doi: 10.3758/s13423-016-1044-3

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/

Rieger, M., and Bart, V. K. E. (2016). Typing style and the use of different sources of information during typing: an investigation using self-reports. *Front. Psychol.* 7:1908. doi: 10.3389/fpsyg.2016.01908

Rumelhart, D. E., and Norman, D. A. (1982). Simulating a skilled typist: a study of skilled cognitive-motor performance. *Cogn. Sci.* 6, 1–36. doi: 10.1207/s15516709cog0601_1

Schmidke, D., Matsuki, K., and Kuperman, V. (2017). Analysis of the time-course of complex word recognition. *J. Exp. Psychol. Learn. Memory Cogn.* 43, 1793–1820 doi: 10.1037/xlm0000411

Seidenberg, M. S., and McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychol. Rev.* 96, 523. doi: 10.1037/0033-295X.96.4.523

Shaffer, L. H. (1975). Control processes in typing. *Q. J. Exp. Psychol.* 27, 419–432. doi: 10.1080/14640747508400502

Spivey, M. (2008). *The Continuity of Mind*. New York, NY: Oxford University Press.

Spivey, M. J., and Dale, R. (2006). Continuous dynamics in real-time cognition. *Curr. Dir. Psychol. Sci.* 15, 207–211. doi: 10.1111/j.1467-8721.2006.00437.x

Spivey, M. J., Grosjean, M., and Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proc. Natl. Acad. Sci. U.S.A.* 102, 10393–10398. doi: 10.1073/pnas.0503903102

van Rij, J., Hendriks, P., van Rijn, H., Baayen, R. H., and Wood, S. N. (2019a). Analyzing the time course of pupillometric data. *Trends Hear. Sci.* 23, 1–22. doi: 10.1177/2331216519832483

van Rij, J., Vaci, N., Wurm, L. H., and Feldman, L. B. (2019b). "Alternative quantitative methods in psycholinguistics: implications for theory and design," in *Word Knowledge and Word Usage: A Cross-Disciplinary Guide to the Mental Lexicon*, eds V. Pirrelli, I. Plag, and W. U. Dressler. Berlin: Mouton de Gruyter.

van Rij, J., Wieling, M., Baayen, R. H., and van Rijn, H. (2017). *itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs*. R package version 2.3. Available online at: https://cran.r-project.org/web/packages/itsadug/index.html

West, L. J., and Sabban, Y. (1982). Hierarchy of stroking habits at the typewriter. *J. Appl. Psychol.* 67, 370–376. doi: 10.1037/0021-9010.67.3.370

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: a tutorial focusing on articulatory differences between L1 and L2 speakers of English. *J. Phon.* 70, 86–116. doi: 10.1016/j.wocn.2018.03.002

Wood, S. N. (2017). *Generalized Additive Models: An Introduction With R, 2nd Edn*. Boca Raton, FL: Chapman and Hall; CRC.

Yamaguchi, M., and Logan, G. D. (2014). Pushing typists back on the learning curve: revealing chunking in skilled typewriting. *J. Exp. Psychol. Hum. Percept. Perform.* 40, 592–612. doi: 10.1037/a0033809

Yarkoni, T., Balota, D., and Melvin Yap, M. (2008). Moving beyond coltheart's N: a new measure of orthographic similarity. *Psychon. Bull. Rev.* 15, 971–979. doi: 10.3758/PBR.15.5.971

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX A

Exp 2 &3	Exp 2 & 3	Exp 2*	Exp 3*
prefix+stem+ suffix	**stem+suffix**	**prefix+ stem (NW)**	**prefix+ stem+ suffix (NW)**
unacceptable	acceptable	unaccept	unacceptance
unadaptable	adaptable	unadapt	unadaptness
inalienable	alienable	inalien	inalienical
inalterable	alterable	inalter	inalterness
unanswerable	answerable	unanswer	unanswerance
disauthorize	authorize	disauthor	disauthorist
unavoidable	avoidable	unavoid	unavoidless
recapitalize	capitalize	recapital	recapitalive
decarbonize	carbonize	decarbon	decarbonist
decentralize	centralize	decentral	decentralion
uncheerful	cheerful	uncheer	uncheerate
unclassical	classical	unclassic	unclassiced
unclimbable	climbable	unclimb	unclimbally
incoherent	coherent	incohere	incoherate
uncomfortable	comfortable	uncomfort	uncomfortness
inconsiderate	considerate	inconsider	inconsiderous
uncritical	critical	uncritic	uncriticer
uncynical	cynical	uncynic	uncynicer
undeadly	deadly	undead	undeader
indifferent	different	indiffer	indifferous
undoubtable	doubtable	undoubt	undoubtness
unearthly	earthly	unearth	unearther
unethical	ethical	unethic	unethicor
inexistent	existent	inexist	inexistery
unfavorable	favorable	unfavor	unfavorless
reformalize	formalize	reformal	reformalous
unfriendly	friendly	unfriend	unfriender
unfruitful	fruitful	unfruit	unfruitity
unhealthily	healthily	unhealth	unhealthity
dehumanize	humanize	dehuman	dehumanous
illogical	logical	illogic	illogicer
demagnetize	magnetize	demagnet	demagnetist
unmatchable	matchable	unmatch	unmatchment
unmechanical	mechanical	unmechanic	unmechanicly
unmetrical	metrical	unmetric	unmetricer
unmindful	mindful	unmind	unmindity
demoralize	moralize	demoral	demoralism
unmusical	musical	unmusic	unmusicer
renormalize	**normalize**	**renormal**	**renormalist**
denuclearize	nuclearize	denuclear	denuclearist
deodorize	odorize	deodor	deodorist
inorganic	organic	inorgan	inorganer
unoriginal	original	unorigin	unorigined
impersonal	personal	imperson	impersoner
depolarize	polarize	depolar	depolarist
depoliticize	politicize	depolitic	depoliticage
unpowerful	powerful	unpower	unpowerism
unprincely	princely	unprince	unprincery
unprofitably	profitably	unprofit	unprofitsome
unsightly	sightly	unsight	unsighten
insolvent	solvent	insolve	insolvery
unspiritual	spiritual	unspirit	unspiriture
insufferable	sufferable	insuffer	insuffersome
intemperate	temperate	intemper	intemperage
intoxicate	toxicate	intoxic	intoxicage
untrustful	trustful	untrust	untrustion
unutterable	utterable	unutter	unutterless
invigorate	vigorate	invigor	invigorish
revitalize	vitalize	revital	revitalous

APPENDIX B

Mean keystroke trajectories aligned on the **offset** of the stem.

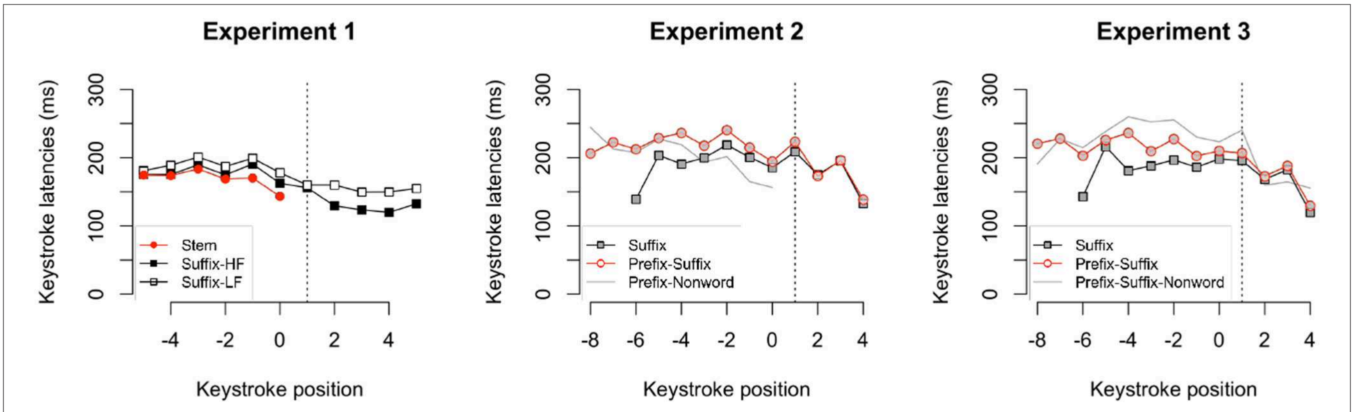


FIGURE B1 | Mean keystroke trajectories (excluding the first keystroke) aligned to the **offset** of the stem for all conditions in Experiments 1, 2, and 3. The onset of the **suffix** is indicated with a dashed vertical line.

frontiers in Communication

ORIGINAL RESEARCH
published: 05 July 2019
doi: 10.3389/fcomm.2019.00029

The Development of Idiom Knowledge Across the Lifespan

Simone A. Sprenger[1]*, *Amélie la Roi*[1] *and Jacolien van Rij*[2]*

[1] Center for Language and Cognition, Faculty of Arts, University of Groningen, Groningen, Netherlands, [2] Department of Artificial Intelligence, Faculty of Science and Engineering, University of Groningen, Groningen, Netherlands

Knowledge of multi-word expressions, such as *break the ice*, is an important aspect of language proficiency that so far we have known surprisingly little about. For example, it is largely unknown how much variability there is between speakers with respect to the number of different items that they know, or what factors contribute to their acquisition. This lack of knowledge seriously limits the generalizability of experimental studies on the production and comprehension of multi-word expressions (usually idioms) and generally suggests that there still is a sizable unknown territory of language knowledge to explore. Here, we present the results of two familiarity ratings for a large sample of Dutch idioms and a large number of participants that varied in age between 12 and 86 years old. The data show considerable variation between participants and between idioms. Non-linear mixed-effects regression analyses revealed that the age of participants, but not their education, as well as the frequency and decomposability of the idioms influenced the familiarity scores. Our findings suggest that the knowledge of multiword expressions develops across the lifespan, is acquired from exposure, and—in participants younger than about 40 years of age—varies with item decomposability.

Keywords: idioms, multiword expressions, vocabulary, aging, decomposability, development, familiarity

Edited by:
Gonia Jarema,
Université de Montréal, Canada

Reviewed by:
Giorgio Arcara,
IRCCS Fondazione Ospedale San Camillo, Italy
Debra Titone,
McGill University, Canada

***Correspondence:**
Simone A. Sprenger
s.a.sprenger@rug.nl
Jacolien van Rij
j.c.van.rij@rug.nl

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *07 November 2018*
Accepted: *04 June 2019*
Published: *05 July 2019*

Citation:
Sprenger SA, la Roi A and van Rij J (2019) The Development of Idiom Knowledge Across the Lifespan. Front. Commun. 4:29. doi: 10.3389/fcomm.2019.00029

INTRODUCTION

In everyday language use, many concepts are expressed by multi-word expressions, such as *hit the road* (depart), *break the ice* (relieve social tension by means of a remark) or *how are you* (a formula exchanged when people meet). These expressions must be learned alongside the words and rules that enable us to generate new sentences and represent an important aspect of what Pawley and Syder (1983) referred to as *nativelike* language proficiency. Based on analyses of conversational data, they estimated the number of such expressions in English as *hundreds of thousands* and suggested that access to these prespecified expressions in long-term memory is a prerequisite for fluent speech. Yet, even though the importance of multi-word expressions has been recognized in psycholinguistics (as evidenced by numerous experimental studies on the acquisition, processing and production of idioms, which we shortly discuss below), our knowledge about these processing units is still very limited. That is, in contrast to our knowledge about single words, we do not know what factors constrain the multi-word vocabulary and the way in which it varies between speakers. Here, we therefore want to explore how speaker characteristics (age, education) and item characteristics (frequency, decomposability) conjointly affect the acquisition of the Dutch idiom vocabulary across the lifespan.

In an exploration of what he called *the boundaries of the lexicon* (and thus the theoretical scope of grammatical theories), Jackendoff (1995) argued that the large number of multiword expressions that speakers of a language know and recognize—which he estimated at about the same size as the number of single words—must in fact be considered entries in the mental lexicon. He illustrated his position with the *wheel of fortune* corpus, which included about 600 compounds, idioms, names and clichés, all considered sufficiently familiar to native speakers to be included in a popular TV game show that required participants to guess these phrases with a few hints. Examples include *I cried my eyes out*, *a breath of fresh air* and *May the Force be with you*.

While the nature of the underlying representations of such well-known phrases is still a matter of debate in linguistics and psycholinguistics (e.g., Cacciari and Tabossi, 1988; Fillmore et al., 1988; Cutting and Bock, 1997; Jackendoff, 1997; Titone and Connine, 1999; Sprenger et al., 2006; Libben and Titone, 2008), most idiom researchers will agree that they need to be included in the mental lexicon. However, our knowledge about this part of the lexicon is still limited. That is, we do not know how many multi-word expressions a speaker can be expected to be familiar with, or what this knowledge depends on. Estimates in the literature (such as Pawley and Syder, 1983, *hundreds of thousands*) are often extrapolations from small samples of conversation. At the same time, collections of multiword expressions in dictionaries or analyses of large corpora (e.g., Moon, 1998) can only provide upper boundaries for the knowledge that a native speaker might acquire. Neither method can provide us with a reliable estimate of the multiword vocabulary, or the conditions that affect its size.

Psycholinguistic approaches to multiword expressions typically focus on idioms. Apart from the fact that they form relatively fixed combinations of words, their meanings are not a direct function of their constituent words, making them an interesting test case for theories of language comprehension and production. For example, depending on the context, the English phrase *to break the ice* either refers to relieving the tension in a social situation or to the actual process of crushing frozen water. However, given a context that fits better with the figurative interpretation, native speakers can easily retrieve the correct form and meaning from memory (in production and comprehension, respectively).

Experimental work that tries to uncover the representations and processes that are responsible for the fast and efficient production and comprehension of idioms depends on high-quality stimulus materials. There are two main criteria that play role in this context: first, the idioms must be representative for a larger collection of items (e.g., with respect to the relationship between form and meaning), and second they must reflect the subjects' knowledge. This second criterion is especially difficult to fulfill. Does every speaker of English know the idiom *to kick the bucket,* or is that knowledge mostly restricted to the subset of idiom researchers? What other well-known expressions are there, and where are these items located in the frequency distribution? In idiom studies, questions about specific items are often answered on an *ad-hoc* basis, with stimulus materials being rated for familiarity in the context of a specific study. The number of items in these studies rarely exceeds twenty (e.g., Bobrow and Bell, 1973: 5 items; Swinney and Cutler, 1979: 22 items; Cacciari and Tabossi, 1988: 20 items; Cutting and Bock, 1997: 36 items; Gibbs, 1991: 20 items; Sprenger et al., 2006: 16 items;), and it is unclear in how far those are representative for the category of idioms as a whole. This lack of knowledge is a fundamental problem for psycholinguistic research on idiom production and comprehension, as it limits the potential generalizability of our data.

For various languages, such as English (Titone and Connine, 1994; Libben and Titone, 2008; Bulkes and Tanner, 2017; Nordmann and Jambazova, 2017), French (Caillies, 2009; Bonin et al., 2013, 2017), German (Citron et al., 2016), Italian (Tabossi et al., 2011) and Chinese (Li et al., 2016), norms have been published with the aim to increase the reliability of stimulus material in psycholinguistic studies on idioms. These norms provide a number of interesting measures, such as familiarity, decomposability, predictability or emotional valence, for several hundreds of items per language. That is, for the average speaker of the language in question, these norms provide a best guess about how a specific item scores on the various dimensions, making it possible for researchers to select items from the corresponding distributions.

However, while clearly increasing the reliability and validity of idiom tasks, the use of norms is not without problems either. It is important to realize that there is no such thing as an average native speaker: they differ with respect to socio-economic backgrounds, education, personality, and age. Given the effect of such variables on the sizes of our vocabularies at large (Brysbaert et al., 2016), it is conceivable that there are considerable individual differences in the idiom vocabulary as well. For example, Brysbaert et al. (2016) showed that the single-word vocabulary expands rapidly during adolescence, but keeps growing steadily until old age, with an average increase of one word per two days. In other words, age has an important effect on vocabulary that exceeds well-beyond the initial stages of language acquisition and cognitive maturation. Yet, idiom norming studies traditionally do not take this factor into account. They usually average across age, often sample from a student population only (e.g., Li et al., 2016), and sometimes do not mention their participants' age at all (e.g., Bulkes and Tanner, 2017). Whether age affects the idiom vocabulary in a similar way as the single-word vocabulary is therefore an open question.

Here, we want to explore the contribution of age to the development of the idiom vocabulary in more detail. If age indeed played an important role in idiom acquisition, this would have important consequences for the design of experiments that are to reveal the psycholinguistic processes and representations involved in the production and comprehension of idioms. Apart from the need to calibrate idiom norms for age, an age effect on idiom knowledge would stress the role of individual differences on online idiom comprehension. Reports on such effects so far have been few, but fairly consistent. Cain et al. (2005), for example, studied the relationship between reading comprehension and idiom interpretation in 9-year olds and found that poor comprehenders were less able to use context when interpreting opaque, but not transparent (or rather, decomposable) idioms. Cacciari et al. (2007) compared slow and fast participants in a comprehension task and found that slow participants needed more perceptual input to identify an idiom and to activate its meaning. Columbus et al. (2015) found effects of executive control capacity on reading times for metaphors, but not for idioms. In contrast, Cacciari et al. (2018) found that idiom comprehension was affected by individual differences in working memory capacity, inhibitory control, and crystallized verbal intelligence, as well as personality-related variables (State Anxiety and Openness to Experience). Taken together, these studies indicate that individual differences affect online idiom comprehension processes, and thus are likely to affect acquisition as well. However, none of the studies considered age as a separate factor.

How would we expect age to affect the idiom vocabulary? First, the pattern that was observed by Brysbaert et al. (2016) for the development of the single-word vocabulary may be further delayed by the late development of figurative competence (i.e., the age at which children are able to understand an idiom's figurative interpretation, at about 9 years of age; Levorato and Cacciari, 1992), as well as by the relatively abstract concepts that are expressed by many idioms. So far, there are only few empirical data to backup this assertion, as developmental research on idioms has mostly focused on figurative competence, rather than the age at which children acquire specific tokens (e.g., Nippold and Martin, 1989; Levorato and Cacciari, 1992; Nippold and Rudzinski, 1993; Nippold and Taylor, 1995; Nippold and Duthie, 2003; Hung and Nippold, 2014). For example, Nippold and Martin (1989) report an increase in the ability to interpret idioms from the age of 14–17. As their observations are based on only twenty items per subject, we cannot draw conclusions about the size of the subjects' idiom vocabularies.

Beyond the age of adolescence, there are likewise only few data points to sketch the acquisition curve. A study by Kuiper et al. (2009) shows a rise in idiom knowledge until the age of 50–60 years, followed by a slight drop-off in the 65+ cohort (ten subjects per cohort). A drawback of this study is that the observations (based on 20 items) are not backed up by inferential statistics, making it difficult to judge their reliability. However, the pattern has partly been confirmed by Escaip (2015). Replicating Kuiper et al.'s (2009) study in Spanish, English, and French, she found a significant positive correlation of age with idiom knowledge in all three languages. That is, the older the participants, the more idioms they knew (with ages ranging between 15 and 83). For English, but not for the other two languages, Escaip also found a significant decrease of knowledge for speakers of 65 years and older.

The second important factor that we want to explore here is idiom frequency. In contrast to age, which is a characteristic of the subjects, frequency is a characteristic of the item itself. Similar to single-word acquisition, it is conceivable that frequency can explain a large part of the variance between idioms. This is supported by the observation that, in the past decade, a considerable number of studies has been published that demonstrate an important role for frequency in the acquisition of multi-word sequences. For example, Bannard and Matthews (2008) showed that children as young as 2 years old are sensitive to the frequency with which specific word combinations occur in child-directed speech: when asked to repeat sequences of words such as a *drink of tea*, they make fewer errors and—by the age of 3—also respond faster to high frequent word combinations than to matched low-frequent combinations. Likewise, Arnon and Snider (2010) demonstrated that adults are sensitive to the frequency of compositional multi-word phrases like *don't have to worry*. Their subjects responded faster in a phrasal decision task when the phrases were more frequent. Similar facilitatory effects for high-frequent items have been observed for language production in adult speakers, both for literal and more idiomatic sequences (e.g., Tremblay and Tucker, 2011; Janssen and Barber, 2012; Arnon and Cohen Priva, 2013; Sprenger and van Rijn, 2013).

The third factor that we include here is idiom decomposability, which was defined as the extent to which the idiom word meanings are related to the figurative meaning of the expression (similar to, for example, Rommers et al., 2013). Similar to frequency, decomposability is a feature of the individual idiom that may affect the ease with which a specific item can be acquired. If an idiom is highly decomposable, knowledge about its individual words may help the learner to deduce the idiom's meaning and/or to remember the item more easily when he or she encounters it again, since the words themselves may act as memory cues. This may explain why the poor comprehenders in the study by Cain et al. (2005) did not have difficulties interpreting decomposable idioms, in contrast to opaque idioms. From studies on online idiom processing, we know that decomposability is a relevant factor. Processing advantages have been reported for decomposable idioms over non-decomposable idioms: for example, with respect to sentence verification latencies (Gibbs et al., 1989) and in a lexical decision task that used idioms as primes for target words that were related to the item's figurative meaning (Caillies and Butcher, 2007). However, the exact nature of the way in which decomposability modulates online processing is still disputed, as its effect is not always facilitatory. Titone and Libben (2014) found late inhibitory effects of decomposability in a cross-modal semantic priming task and Titone et al. (2019) observed late inhibitory effects of decomposability during idiom reading. Interestingly, Westbury and Titone (2011) found an interaction of decomposability with age: in a literality judgment task, older adults were relatively slower than younger adults to accept non-decomposable idioms with a literal meaning and made more errors.

In the present article, we want to study the effect of age as an easy to assess speaker characteristic on idiom familiarity and compare it to the effects of idiom frequency and decomposability. If idioms indeed have their own entries in the mental lexicon, the idiom familiarity curve should be highly similar to that for single-word vocabulary (across speakers and items). That is, it should be modulated by age and education, with an early phase of rapid expansion, followed by a long phase of moderate but steady increase, and possibly decrease (as in Kuiper et al., 2009). Per item, this effect should be modulated by frequency, as we can expect the probability of acquisition to be a function of exposure. It may also be affected by idiom decomposability, which is supposed to reflect the ease with which an item can be analyzed, encoded, and retrieved (Caillies and Butcher, 2007). To test these predictions, we collected familiarity ratings for

194 Dutch idioms in two online rating studies and assessed the corresponding corpus frequencies. In addition to the ratings, respondents provided information about their gender, age, and level of education.

THE IDIOM DATABASE

For the exploration of the effect of age on idiom familiarity (Study 1 and 2 presented below), we have composed a small database with Dutch idioms. The database is available in the supplementary materials[1] and contains 189 Dutch idioms with their meaning and associated frequency counts. For all idioms (and control items, as explained below) we additionally collected decomposability ratings in an online questionnaire.

Materials and Methods

Materials

Ninety-nine Dutch idioms with two nouns were collected for Study 1. They were not controlled for syntactic structure or position of the nouns, but often contained prepositional phrases. The number of nouns was controlled with respect to the item's usability in an unrelated behavioral experiment. In addition to the experimental items, four German idioms were literally translated to Dutch and included as control items. All items were presented in past tense and preceded by the temporal adverb "Toen": (at a time in the past), for example "Toen kwam de aap uit de mouw." (*Then the monkey came out of the sleeve*, which means that the true nature of a situation, the true character of a person, or a hidden motive was being revealed).

Ninety Dutch idioms with one noun were collected for Study 2. Again, syntactic structure or noun position were not controlled for. All items were presented in past tense and preceded by the temporal adverb "Toen" (at a time in the past), for example "Toen zette hij hem op straat." (*Then he put him on the street*, which means *then he laid him off*) In contrast with Study 1, no control items were included. Thus, all idioms were existing Dutch idioms.

Frequencies for the idioms and translated German idioms were obtained from the Lassy Large corpus (Van Noord et al., 2013), a 700-million-word corpus of Dutch texts with automatically assigned syntactic annotations that is combined of both spoken and written sub-corpora (including the Dutch *Wikipedia*). By searching for lemmas, rather than exact word matching, most idioms were detected: the counts ranged between 0 (4 items) and 4,688. Surprisingly, three of the five control also were found in the corpus, probably due to their similarity with other Dutch idioms (for example, the German idiom *Then he shot sparrows with cannons* is very similar to the Dutch idiom *Then he shot mosquitos with cannons*). Before analysis, the frequency counts were log-transformed. **Figure 1** shows the distribution of the log-transformed frequencies.

Participants

The decomposability questionnaire was advertised under students of the University of Groningen. We restricted the age range to 18–25 years old, to keep the decomposability ratings consistent with earlier studies (Rommers et al., 2013). The data consisted of 57 entries, but we excluded one participant who was not monolingual Dutch (a Frisian-Dutch bilingual), 21 participants who contributed less than ten ratings, and one participant whose age did not match the target age range. The clean data consisted of 34 participants in the age range 21–26 years old (mean 24.3 years old; 8 men) who contributed each 15–98 ratings (mean 89.9). Participants did not receive compensation for their participation.

Procedure

The questionnaire was implemented using the survey software Qualtrics (Qualtrics, Provo, UT). Participants could anonymously access the questionnaire with a link. At the start of the experiment, participants were informed on the goal of the survey and gave their consent that their participation was voluntary. Participants were asked to read idioms and to judge to what extent the meaning of the individual words was related to the figurative meaning of the expression as a whole (cf. Rommers et al., 2013). They had to click on one of five radio buttons, labeled from left to right as "1 (geen relatie tussen individuele woorden en figuurlijke betekenis)" (*no relation between the individual words and the figurative meaning*), "2," "3," "4," and "5 (sterke relatie tussen individuele woorden en figuurlijke betekenis)" (*strong relation between the individual words and the figurative meaning*), or on a sixth radio button labeled as "ik ben niet bekend met deze uitdrukking" (*I am not familiar with this idiom*). Three idioms were presented individually at the start of the questionnaire to serve as anchors for the range of the decomposability scale (*anchoring*), but later idioms were presented in a random order. The idioms were divided in two lists of each 100 items (including the anchors). Each participant saw only one of the two lists.

Analyses

The data were analyzed using Generalized Additive Mixed Models (Hastie and Tibshirani, 1990; Wood, 2017; GAMMs), a *non-linear* mixed-effects regression method. GAMMs do not assume a linear relationship between the dependent variable and a covariate, but the relationship is estimated using penalized regression splines. The method does not require the user to specify the shape of the regression line on beforehand, but it is estimated based on the data. Other reasons for choosing this non-linear regression method are that it allows to include tensor product interactions for estimating interactions between multiple non-linear covariates, and it allows to include non-linear random effects (see for introductions Wieling, 2018; van Rij et al., in press). The statistical analyses are performed in R version 3.4.4 (2018-03-15; R Core Team, 2018), using the package *mgcv* version 1.8-24 (Wood, 2017) implementing GAMMs, and the package *itsadug* 2.3 (van Rij et al., 2017) for evaluation and visualization of the statistical models.

[1]Supplementary Materials are available at https://git.lwp.rug.nl/p251653/development-idiom-knowledge

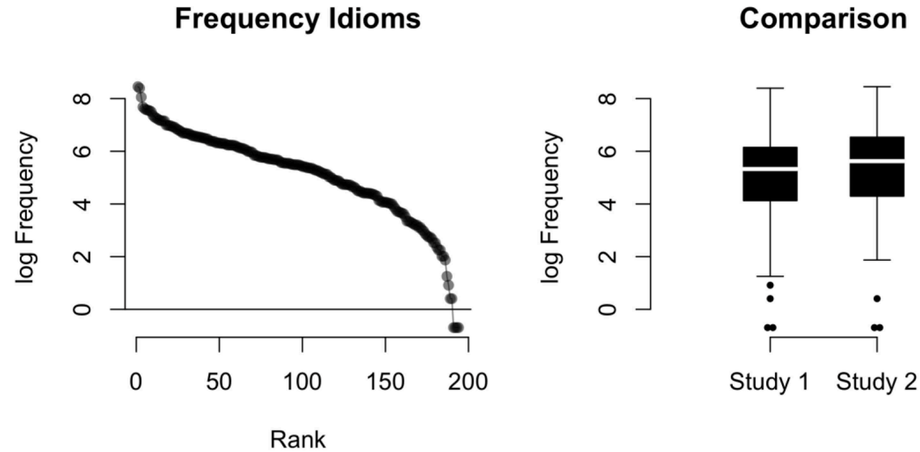


FIGURE 1 | Left: Frequency (log-transformed) of the 194 Dutch idioms and translated German idioms, ordered. **Right:** Comparison of the frequency distributions in Study 1 and Study 2.

Decomposability Ratings

From the 3,056 responses, 504 (16.5%) were of the category *I am not familiar with this idiom* (henceforth "unfamiliar" responses). These responses were excluded from the analysis. A logistic mixed-effects regression analysis revealed that the proportion of "unfamiliar" responses was significantly influenced by the idioms' frequencies [$\chi^2_{(2)} = 24.02$, $p < 0.001$]: the proportion of "unfamiliar" responses is larger for low-frequent idioms than for high-frequent idioms (see Supplementary Materials for the complete analysis).

All idioms were seen by at least thirteen participants. However, the number of actual decomposability ratings (i.e., when participants did *not* give an "unfamiliar" response) varied strongly between idioms, ranging from 2 to 34 (mean 13.1). **Figure 2** shows this variation in the number or ratings that was collected for each idiom: On the right end of the x-axis, there is one idiom that received a decomposability score from all 34 participants, because it was included as anchor. At the left end of the x-axis we find one of the translated German idioms with no decomposability ratings. All 13 participants who were presented with this idiom indicated that they were not familiar with it. We excluded this item from further analysis, accordingly.

We did not use the average rating per idiom as decomposability score, to avoid a potential subject bias influencing the decomposability scores for the idioms with a low number of ratings. Instead, we fitted a GAMM with random effects for participants and idioms to account for the participants' response biases and the variation between idioms. Random effects allow for partial pooling: the estimates for idioms that only have a few observations will pull toward the average (shrinkage); and the idiom estimates may be corrected for subject biases, as the subject mean is taken into account. From this statistical model we extracted an estimated decomposability score for each idiom (the script is available in the Supplementary Materials). To fit the ordered categorical nature of the decomposability ratings (5-point scale), we used the GAM ordered categorical family (Wood et al., 2016). **Figure 3** (left panel) visualizes the difference between the mean rating scores (x-axis) and the estimated decomposability scores (y-axis). **Figure 3** (right panel) compares the estimated decomposability scores for Study 1 and Study 2.

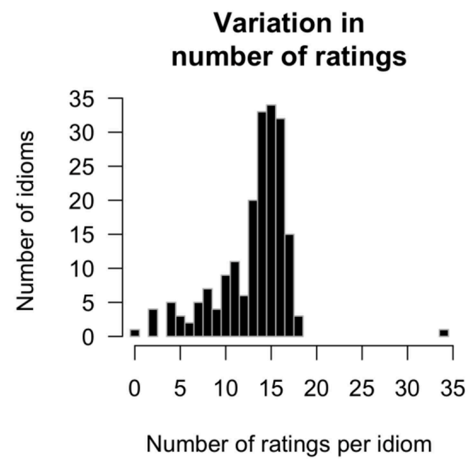


FIGURE 2 | Histogram of the number of decomposability ratings per idiom. The x-axis shows the number of decomposability ratings, and the y-axis the number of idioms that received that number of ratings.

Finally, we analyzed the effects of the idiom's frequency on the decomposability score. We used the GAM ordered categorical family (Wood et al., 2016) to fit the decomposability ratings (5-point scale). The log-transformed frequencies were included as non-linear main effect. In addition, by-Subject non-linear random smooths were included for Frequency and random intercepts for Idiom. However, the effect of Frequency was not significant [$F_{(1.001,2381.614)} = 2.69$; $p = 0.1$].

In the following sections we will use item frequencies and decomposability ratings as predictors for the familiarity ratings of Study 1 and Study 2.

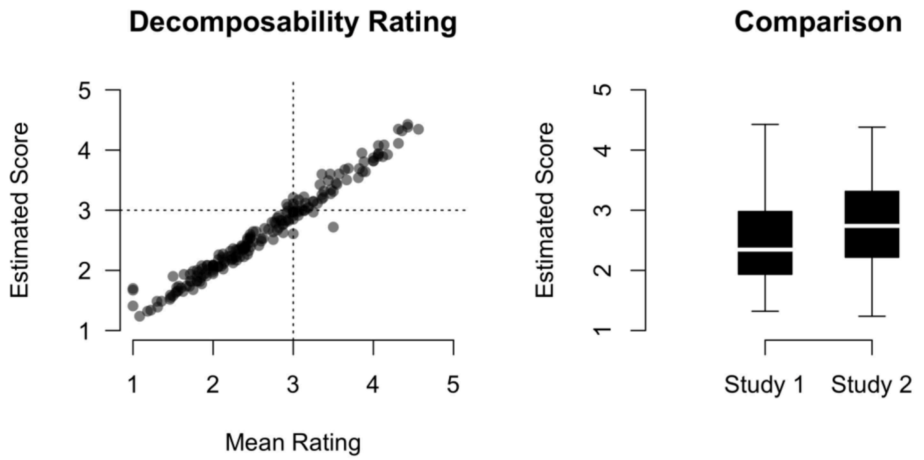


FIGURE 3 | Decomposability scores. **Left:** Mean rating scores (x-axis) vs. estimated rating scores with partial pooling (y-axis). The dashed lines mark the center of the scale. **Right:** Comparison of the estimated rating scores in Study 1 and Study 2.

STUDY 1: TWO-NOUN IDIOMS

In the first online questionnaire, we collected familiarity ratings for 104 Dutch idioms and control items with two nouns.

Materials and Methods

Participants

The questionnaire was advertised via social media (Facebook and Whatsapp) in the personal networks of the first and last author. The data consisted of 319 entries, but we excluded 25 participants who were not monolingual Dutch (13 of which were Frisian-Dutch bilingual). Subsequently, 37 participants were removed because the participants contributed less then ten ratings. The clean data consisted of 257 participants in the age range 12–86 years old (mean 37.7; 65 men) who contributed 96–104 ratings. Participants did not receive compensation for their participation.

Materials and Design

Ninety-nine Dutch idioms with two nouns were collected for this study. In addition, five German idioms were literally translated to Dutch and included as control items. The form of the materials *(Then the monkey came out of the sleeve)* was identical to the one described in section The idiom database.

All participants saw the same 104 idioms, but the order or presentation was randomized per participant. In addition to rating the idioms, participants were asked background questions about their gender, the year and month of birth, and their highest completed education (elementary school, high school, vocational education, or university).

Procedure

Participants could perform the questionnaire online on their computer, laptop, or tablet, or smartphone. The type of device was not registered. We have implemented the questionnaire using the survey software Qualtrics. Participants could anonymously access the questionnaire with a link. At the start of the experiment, participants were informed on the goal of the survey and they gave their consent that their participation was voluntary.

Participants were asked to read idioms and to judge whether their age peers would recognize this idiom when it would be used in a talk show. They had to click on one of five radio buttons, labeled from left to right as "1 (nog nooit gehoord)" (*never heard before*), "2," "3," "4," and "5 (heel bekend)" (*very well-known*). Three idioms were presented individually at the start of the questionnaire to serve as anchors for the range of the familiarity scale (*anchoring*), but later idioms were presented all at once in a long list in a random order to reduce the number of mouse clicks.

Results

Figure 4 shows the average rating per participant, plotted against their age (Left panel), the average rating per idiom and participant age (Center panel), and the average rating per age, collapsed over participants and idioms (Right panel). What immediately stands out from these plots is the variation between participants and between items. A closer look reveals that with younger ages the variation is larger than with older ages. Finally, the grand averages show us a clear increase in idiom familiarity over age, which continues in older ages. The ratings for each education level and the average age per education level are presented in **Table 1**.

The data were analyzed using GAMMs (Hastie and Tibshirani, 1990; Wood, 2017). We included *Education* and *Gender* as categorical predictors in the statistical model. Education is a three-level predictor describing the participant's education using the categories "University," "Vocational education," and "Other" (collapsing elementary school and high school). Further, we included the covariates *Age*, the participant's age in years, *Frequency*, the log-transformed frequency of the idiom, *Decomposability*, the estimated decomposability scores, and their interactions, and by-participant random smooths over Frequency and over Decomposability, and by-idiom random smooths over Age. These three random smooths account for variations between

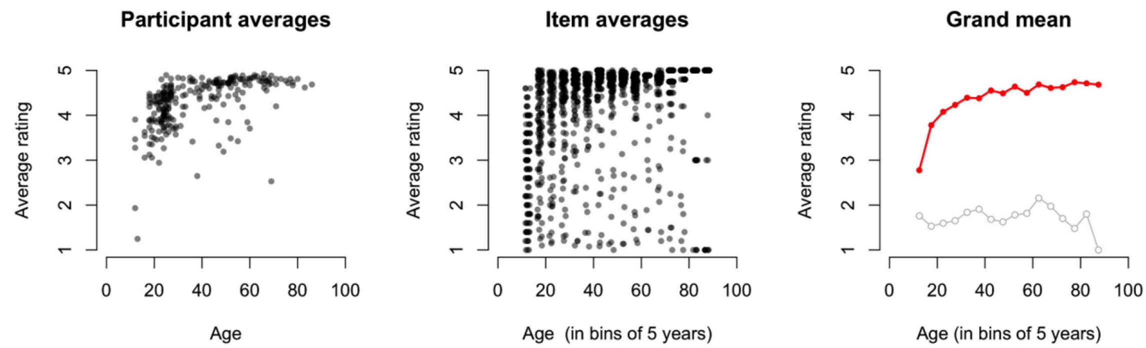


FIGURE 4 | Average familiarity ratings in study 1. **Left:** Average familiarity ratings per participant (y-axis) and the participant's age (x-axis). **Center:** Average familiarity ratings per idiom (y-axis) per age group (x-axis; in bins of 5 years, and jitter added for presentation purposes). **Right:** Grand average of familiarity ratings (y-axis) per age group (x-axis; in bins of 5 years). The solid red dots plot the grand average collapsed over all idioms, and the gray circles plot the grand average of the control items only.

TABLE 1 | Average ratings and average ages per education level in Study 1.

Education	N	Gender	Age		Rating
		Women	Mean	Median	
Elementary school	7	4	13.6	12	2.9
High school	39	31	31.8	21	4.2
Vocational education	35	27	50.7	51	4.6
University	176	130	37.4	29	4.3

participants and idioms, capturing random intercept differences and non-linear random deviations from the regression lines.

The dependent variable is the rating that participants provided for each idiom on a five-point rating scale (1 being unknown, and 5 being well-known). To account for the non-Gaussian nature of the dependent variable, the model was fitted using the GAM ordered categorical family, which implements regression for ordered categorical data (Wood et al., 2016). The smoothing parameter estimation method fREML was used, because the number of idioms and participants was too large for using ML. The method fREML allows for discretizing covariates and thereby decreasing the processing time enormously. A disadvantage of using fREML is that model comparisons are less reliable (see Wieling, 2018). Therefore, we did not only test a backward-fitting model comparison procedure using AIC and fREML, but also used the summary statistics and visual inspection of the model to determine the best-fitting model (cf. van Rij et al., 2019).

The manual backward-fitting model comparison procedure suggested that the non-linear three-way interaction between Age, Decomposability, and Frequency was not significantly contributing to the model [$\chi^2_{(4)} = 3.37$, $p > 0.1$; ΔAIC $= -0.66$]. The summary statistics of the full model confirmed that the interaction surface was not significantly different from zero ($F_{(1.0,25792.4)} = 2.16$; $p > 0.1$). The non-linear two-way interaction between Decomposability and Frequency also did not show significance in the model comparison procedure [$\chi^2_{(3)} = 2.06$, $p > 0.1$; ΔAIC $= -0.83$; summary statistics: $F_{(3.2,25792.4)} = 1.21$; $p > 0.1$]. The interaction between Age and Frequency was significantly contributing to the model [$\chi^2_{(3)} = 12.21$, $p < 0.001$; ΔAIC $= -7.10$; summary statistics: $F_{(7.0,25792.4)} = 4.01$; $p < 0.001$]. The interaction between Age and Decomposability was found marginally significant [$\chi^2_{(3)} = 4.11$, $p = 0.42$; ΔAIC $= 1.05$] in the model comparison procedure, but the summary statistics indicated that the interaction surface was different from zero [$F_{(1.0,25792.4)} = 7.48$; $p < 0.01$]. The categorical predictors Gender [$\chi^2_{(1)} = -0.44$, $p > 0.1$; ΔAIC $= -0.02$] and Education [$\chi^2_{(2)} = 0.72$, $p > 0.1$; ΔAIC $= 0.47$] did not contribute to the model, and were excluded from the model. The best-fitting model included the non-linear interactions between Age and Frequency and between Age and Decomposability and the non-linear main effects of Age, Frequency, and Decomposability. The best-fitting GAMM model: Rating ∼ s(Age) + s(LogFreq) + s(Decomp) + ti(Age, LogFreq) + ti(Age, Decomp) + s(LogFreq, Subject, bs = 'fs', m = 1) + s(Decomp, Subject, bs = 'fs', m = 1) + s(Age, Sentence, bs = 'fs', m = 1), with the last three terms being non-linear random effects. In the best-fitting model, the main effects of Age [$F_{(3.2,25792.7)} = 31.30$; $p < 0.001$) and Frequency [$F_{(3.192,25792.736)} = 24.39$; $p < 0.001$] were significantly different from zero, but not the main effect of Decomposability [$F_{(1.1,25792.7)} = 3.56$; $p = 0.068$].

Figure 5 visualizes the estimates of the best-fitting GAMM by plotting the *fitted effects* (i.e., the sum of all model terms, which results in the model's estimate of the familiarity rating). The left panel shows the estimated effect of Age on the familiarity rating: the familiarity increases with age until around 60 years. Note that the values of the fitted effects are not directly comparable with the rating scale, because ordered categorical GAMMs use transformed values. The estimated cut-off points are added in red and these indicate how the transformed values relate to the response ratings. The Center panel shows the interaction between Age and Log Frequency in a contour plot, with on the z-axis the model's estimates for the familiarity ratings, again on the transformed scale. The interaction surface shows that for medium and high frequency values, the familiarity increases with age and is at ceiling for older participants.

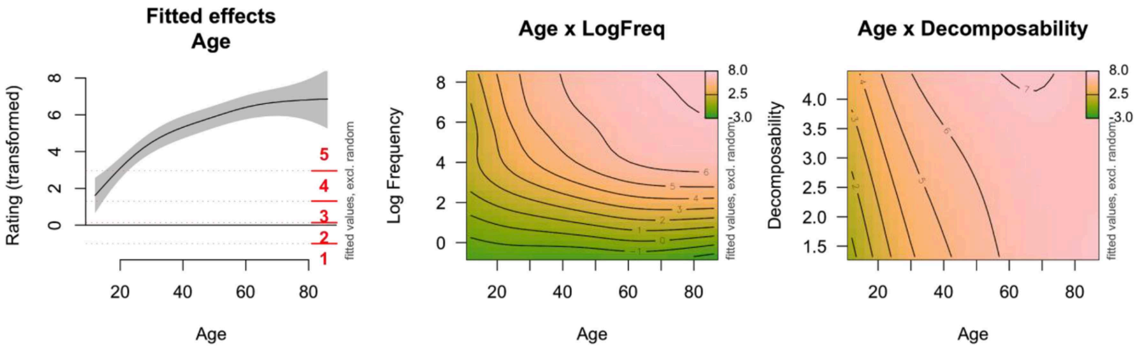


FIGURE 5 | Fitted effects of GAMM analysis of Study 1. Random effects are set to zero. **Left:** Estimated effect of Age (with pointwise 95% confidence intervals) for median values of LogFrequency (5.35) and Decomposability (2.35). Age is presented on x-axis, and the estimated familiarity rating on the y-axis. The red horizontal lines indicate the estimated cut-off points that relate the estimated values to the five response options. **Center:** Non-linear interaction between Age (x-axis) and Log Frequency (y-axis). The colors and contour lines indicate the value of the estimated ratings (z-axis). **Right:** Non-linear interaction between Age (x-axis) and Decomposability (y-axis). The colors and contour lines indicate the value of the estimated ratings (z-axis).

However, for the lowest frequency values, all age groups respond with a low familiarity value (i.e., the horizontal lines at the bottom). This is probably caused by the low frequency and control items, which also were rated as unfamiliar by the older participants. The Right panel visualizes the interaction between Decomposability and Age. Idioms with low decomposability scores are rated lower in familiarity than idioms with high decomposability scores. However, this decomposability effect is only found for younger and middle-aged adults, not for the older adults (> 60 years).

STUDY 2: ONE-NOUN IDIOMS

To verify whether the age effect also applies to other idioms and participants, we ran a second online questionnaire in which familiarity ratings for Dutch idioms were collected. This time the idioms had a different structure: instead of two nouns, the majority of these idioms contained one noun. The procedure of the experiment was exactly the same, only the participants and materials were different.

Materials and Methods

Participants

The questionnaire was advertised via social media (Facebook and Whatsapp) in the personal networks of the second author. The data consisted of 173 entries, but we excluded 56 participants who were not monolingual Dutch (48 of which were Frisian-Dutch bilingual). Subsequently, 12 entries were removed because the participants contributed less than ten ratings. The clean data consisted of data from 105 participants in the age range 19–76 years old (mean 42.9; 20 men) who contributed 15–90 ratings. Participants did not receive compensation for their participation.

Materials and Design

Ninety Dutch idioms with one noun were collected for this study. All items were presented in past tense and preceded by the temporal adverb "Toen" (at a time in the past), for example "Toen zette hij hem op straat." (*Then he put him on the street*, then he laid him off) In contrast to Study 1, no control items were included. Thus, all idioms were existing Dutch idioms.

All participants saw the same 90 idioms, but the order of presentation was randomized per participant. In addition to rating the idioms, participants were asked background questions about their gender, the year and month of birth, and their highest completed education (elementary school, high school, vocational education, or university).

Results

Figure 6 shows the average familiarity rating over age for participants (left panel), for idioms (center panel), and the grand average, collapsed over participants and idioms (right panel). Again, we see a large variation between participants and between items, maybe even more than in the data of Study 1 (**Figure 4**). The right panel shows the average rating per age, collapsed over participants and idioms. The plot shows an increase in average rating with age until the age of 55, after which the average ratings decrease again. This decrease was not visible in the averages of the data from Study 1. The ratings for each education level and the average age per education level are presented in **Table 2**.

The data of Study 2 were analyzed in the same way as Study 1, using Generalized Additive Mixed Models (Hastie and Tibshirani, 1990; Wood, 2017; GAMMs). We included *Education* and *Gender* as categorical predictors in the statistical model. Education is a three-level predictor describing the participant's education using the categories "University," "Vocational education," and "Other" (collapsing elementary school and high school). Further, we included the covariates *Age*, the participant's age in years, *Frequency*, the log-transformed frequency of the idiom, *Decomposability*, the decomposability scores, and we included as random effects by-participant random smooths over Frequency, by-participant random smooths over Decomposability, and by-idiom random smooths over Age.

The dependent variable is the ratings that participants provided for each idiom on a five-point rating scale (1 being

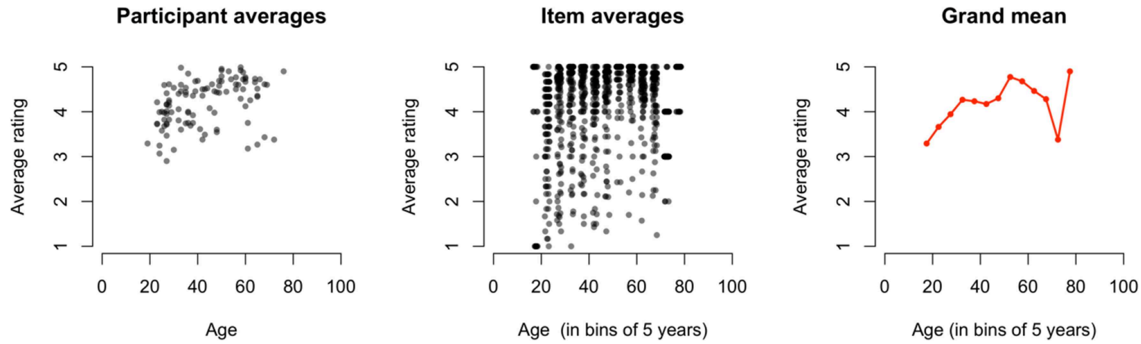


FIGURE 6 | Average familiarity ratings in study 2. **Left:** Average familiarity ratings per participant (y-axis) in relation to and the participant's age (x-axis). **Center:** Average familiarity ratings per idiom (y-axis) per age group (x-axis; in bins of 5 years, and jitter added for presentation purposes). **Right:** Grand average of familiarity ratings (y-axis) per age group (x-axis; in bins of 5 years).

TABLE 2 | Average ratings and average ages per education level in Study 2.

Education	*N*	Gender	Age		Rating
		Women	Mean	Median	
High school	7	7	52.0	64	4.1
Vocational education	28	24	45.3	47	4.1
University	70	54	41.1	37.5	4.2

unknown, and 5 being well-known). To account for the non-Gaussian nature of the dependent variable, the model was fitted using the GAMM ordered categorical family, which implements regression for ordered categorical data. As before, the smoothing parameter estimation method fREML was used, because the number of idioms and participants was too large for using ML.

The manual backward-fitting model comparison procedure suggested that the non-linear three-way interaction between Age, Decomposability, and Frequency was not significantly contributing to the model [$\chi^2_{(4)} = 3.67$, $p > 0.1$; ΔAIC $= -0.44$; summary statistics: $F_{(6.4,8111.0)} = 1.51$; $p > 0.1$]. The non-linear two-way interaction between Decomposability and Frequency also did not show significance in the model comparison procedure [$\chi^2_{(3)} = 3.11$, $p > 0.1$; ΔAIC $= 0.19$; summary statistics: $F_{(1.0,8111.0)} = 2.71$; $p = 0.1$]. The interaction between Age and Frequency was significantly contributing to the model [$\chi^2_{(3)} = 5.23$, $p = 0.015$; ΔAIC $= 2.69$; summary statistics: $F_{(5.9,8111.0)} = 2.41$; $p = 0.018$], and also the interaction between Age and Decomposability [$\chi^2_{(3)} = 6.25$, $p = 0.006$; ΔAIC $= -3.15$; summary statistics: $F_{(2.07,8111.03)} = 6.85$; $p < 0.001$]. The categorical predictors Gender [$\chi^2_{(1)} = 1.11$, $p > 0.1$; ΔAIC $= 0.14$] and Education [$\chi^2_{(2)} = 1.30$, $p > 0.1$; ΔAIC $= -0.35$] did not contribute to the model, and were excluded from the model. The best-fitting model included the non-linear interactions between Age and Frequency and between Age and Decomposability and the non-linear main effects of Age, Frequency, and Decomposability. As a result, we ended with the same specification for the best-fitting GAMM model as in the analysis of Study 1: Rating ~ s(Age) + s(LogFreq) + s(Decomp) + ti(Age, LogFreq) + ti(Age, Decomp) + s(LogFreq, Subject, bs = 'fs', m = 1) + s(Decomp, Subject, bs = 'fs', m = 1) + s(Age, Sentence, bs = 'fs', m = 1), with the last three terms being non-linear random effects. The main effects of Age [$F_{(3.1,8106.6)} = 6.62$; $p < 0.001$] and Frequency [$F_{(2.5,8106.6)} = 17.3$; $p < 0.001$] were significantly different from zero, but not the main effect of Decomposability [$F_{(1.0,8106.6)} = 0.09$; $p > 0.1$].

Figure 7 illustrates the fitted effects estimates of the GAMM analysis of the familiarity ratings of Study 2. The main effects regression line for Age indicates that the ratings increase with age until age 55, and decrease a little for the oldest participants. The oldest participants also show largest uncertainty around the estimates, because there are not many participants around 70. The center panel of **Figure 7** shows the interaction between Age and Frequency: idioms with a lower frequency result in lower familiarity ratings than idioms with a higher frequency, but this effect is stronger for young participants. The right panel of **Figure 7** shows the interaction between Age and Decomposability. The plot suggests that the decomposability scores influence the familiarity ratings of younger participants (< 40 years old), but not of older participants.

COMPARISON OF STUDY 1 AND STUDY 2

The findings reported both in Study 1 and Study 2 suggest that the familiarity of idioms increases with age, idiom frequency, and decomposability score. To test whether the trends for age, frequency, and decomposability are the same in the two experiments, we compared the estimated effects of the best-fitting models. As we used ordered categorical GAMMs, we cannot compare the model estimates directly. Ordered categorical GAMMs model the effects on a continuous scale and estimate the cut-off points that define the boundaries between the categories on the rating scale. These cut-off points are different for the analysis of Study 1 (−1, 0.14, 1.31, 2.96) and Study 2 (−1, 0.06, 1.13, 2.47). Instead, we can extract from the model the probability of the ordered categorical variable being of the corresponding category, and compare these probabilities (Wood et al., 2016). **Figure 8** illustrates the relation between the fitted estimates over

FIGURE 7 | Fitted effects of GAMM analysis of Study 2. Random effects are set to zero. **Left:** Estimated effect of Age (with pointwise 95% confidence intervals) for median values of LogFrequency (5.35) and Decomposability (2.35). Age is presented on x-axis, and the estimated familiarity rating on the y-axis. The red horizontal lines indicate the estimated the cut-off points that relate the estimated values to the five response options. **Center:** Non-linear interaction between Age (x-axis) and Log Frequency (y-axis). The colors and contour lines indicate the value of the estimated ratings (z-axis). **Right:** Non-linear interaction between Age (x-axis) and Decomposability (y-axis). The colors and contour lines indicate the value of the estimated ratings (z-axis).

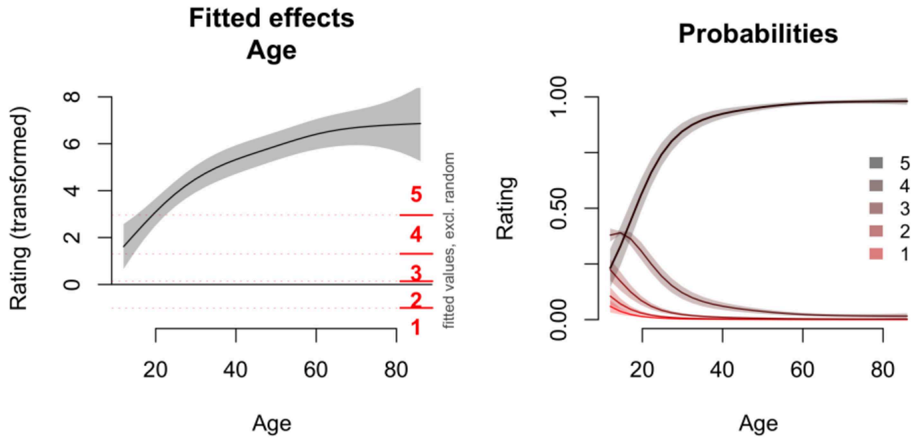


FIGURE 8 | The relation between fitted effects excluding random effects **(Left)**, and the probabilities for the dependent variable being in one of the five categories **(Right)**. The left panel includes the estimated cut-off points in red.

Age (summing over all predictors, including the intercept; Left panel), and the probabilities for the dependent variable being in one of the five categories, using the effect of Age in Study 1 (Right panel).

In the left panel of **Figure 9** we (visually) compared the effect of Age on the probabilities for the response variable being rating 4 or 5. To facilitate the comparison we did not include the other three ratings in the plot. The solid lines are the estimated probabilities based on the best-fitting model fitted on the data of Study 1, whereas the dashed red lines are the estimated probabilities based on the best-fitting model fitted on the data of Study 2. Over all ages, the probability of selecting 5 (very well-known) is higher in Study 1 than in Study 2, but the probability of selecting 4 is higher in Study 2 than in Study 1. Thus, irrespective of Age, the idioms in Study 2 are rated as less familiar than the idioms in Study 1. In addition, we see a clear decrease in selecting response option 5 for older adults (> 60 years) in Study 2, but not in Study 1. Important to mention is that these fitted effects are calculated for a median log-frequency (5.48) and a median decomposability score (2.50).

The center panels of **Figure 9** show the estimated probabilities of the response variable being rating 5 for the data of Study 1 (top) and Study 2 (bottom), and how this probability is influenced by Frequency and Age. The contour plot of Study 1 indicates that low frequency idioms (such as the control items, which were translated German idioms that do not exist in Dutch) are unfamiliar for all age groups, as indicated by the green color which is associated with low probabilities and the horizontal contour lines (i.e., no changes in Age, only in frequency). The high frequency idioms on the other hand are rated as being highly familiar by participants older than 30, as indicated by the pink color which marks a probability of 1. These same high frequency idioms show a sharp increase over age in the probability of being rated as 5 for participants under 30 years, as indicated by the vertical contour lines (i.e., no change in frequency, but in age). The contour plot of Study 2 roughly shows similar patterns (increase in probability with Age and Frequency), but

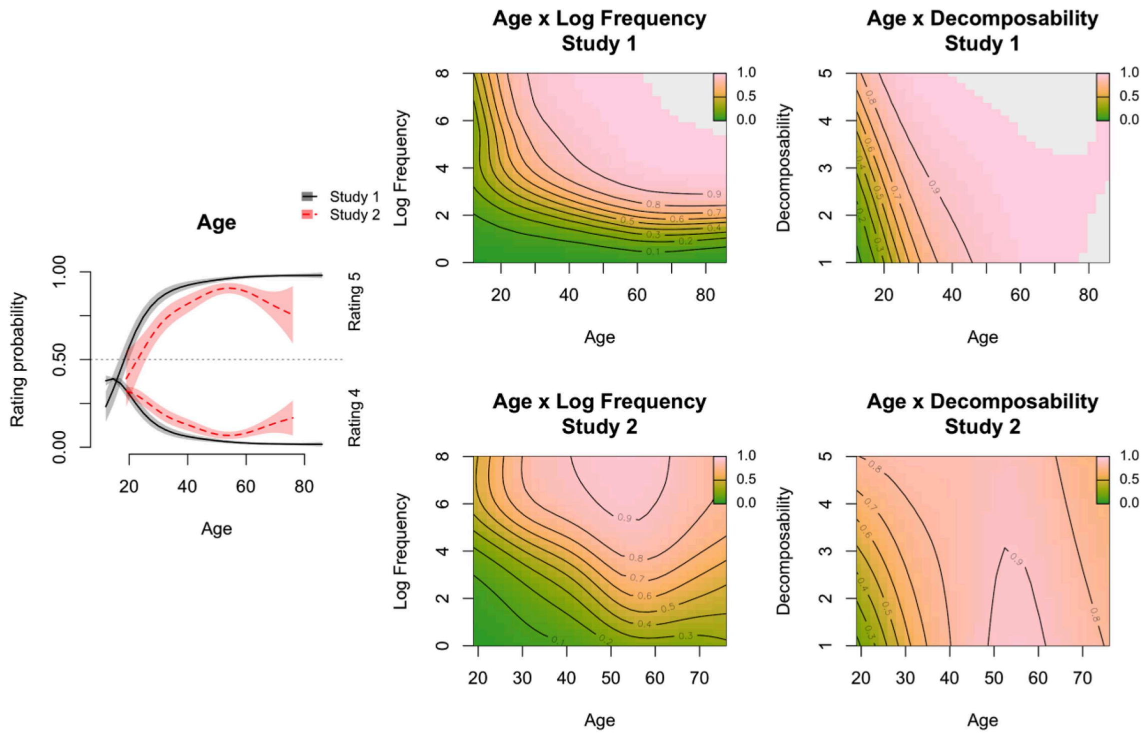


FIGURE 9 | Probabilities of the response variable being rating 5 (highly familiar). The probabilities for Studies 1 and 2 are derived from two different statistical models. Random effects are set to 0. **Left:** Comparing the probabilities of rating 4 and 5 (y-axis) over Age (x-axis) for Studies 1 and 2, with median values for Log Frequency (5.48) and Decomposability (2.50). **Center:** Effects of Age (x-axis) and Log Frequency (y-axis) on the probability of rating 5 (z-axis) for Study 1 (top) and for Study 2 (bottom) with Decomposability set to a median value (2.50). **Right:** Effects of Age (x-axis) and Decomposability (y-axis) on the probability of rating 5 (z-axis) for Study 1 (top) and for Study 2 (bottom) with Log Frequency set to a median value (5.48). Note that the age range is smaller for Study 2 than for Study 1.

the pattern looks more variable. One of the causes of the more variable pattern in Study 2 may be that no control items had been included.

The right panels of **Figure 9** shows the estimated probabilities of the response variable being rating 5 for the data of Study 1 (top) and Study 2 (bottom), and how this probability is influenced by Decomposability and Age. The two contour plots show a very similar pattern: idioms that are perceived as highly decomposable, have a higher probability of being rated with a 5 (highly familiar) by young and old participants. Idioms that are perceived as less decomposable, have a lower probability of being rated with a 5 by younger participants (< 40 years old). On the basis of visual inspection, it seems that the two studies show a stronger difference in the interaction between Age and Frequency than in the effect of Decomposability.

DISCUSSION

We explored the variability in idiom knowledge in a large sample of native speakers of Dutch, divided across two separate idiom familiarity studies. Based on findings for single-word vocabulary, and driven by the assumption that idioms (and other multi-word expressions) can be considered entries in the mental lexicon, we expected to find a familiarity curve that shows an early phase of rapid expansion, followed by a long phase of moderate but steady increase across the lifespan. This pattern has been confirmed in both studies. The grand averages show a clear increase in idiom familiarity over age that proceeds until at least 55 years of age.

We also observed a predicted delay in the rise of the idiom vocabulary, compared to single-word acquisition, by about 10 years, as the steep increase in idiom acquisition levels off after 30 years.

The second main factor of interest was idiom frequency. As with single words, the simple rationale is that the more often speakers come across a specific item, the higher the probability of long-term retention. We therefore expected higher ratings for high-frequent idioms. This pattern has been confirmed in both studies, with frequency significantly impacting on the probability of a high familiarity score. While the low frequency items consistently score low in all age groups, the high frequency idioms are rated as being highly familiar by participants older than 30 and show a sharp familiarity increase over age for participants younger than 30 years. Based on Brysbaert et al.'s (2016) findings for the single-word vocabulary, we further expected to find evidence for an effect of education. However, no such effect was found.

As a third factor, we included independent ratings of idiom decomposability, as they might inform us about the way in which item characteristics affect the ease with which idioms are acquired across age. We indeed find that idioms with low decomposability scores are rated as less familiar than items with high decomposability scores, in both studies. However, this effect seems restricted to the younger raters (< 40 years old). This suggests that the degree to which the individual words are perceived to contribute to the meaning of the idiom as a whole affects the item's learning trajectory. Decomposable idioms may be more easily acquired than non-decomposable idioms, which could explain why Cain et al. (2005) found that children with poor reading comprehension skills had difficulties interpreting non-decomposable, but not decomposable idioms. Whereas the meaning of decomposable idioms can be derived from the meanings of the idiom constituents, the meaning on non-decomposable idioms has to be learned explicitly. Yet, our findings also suggest that once the item has been acquired, the degree to which it is decomposable no longer affects its perceived familiarity. In this context, it is noteworthy that we deliberately limited the age range of the participants who provided the decomposability ratings (18–25 years), to avoid a possible confound of the ratings with age. An interaction of item decomposability and age has been reported for online processing (Westbury and Titone, 2011). In a follow-up on our study on idiom knowledge, it would be interesting to see to what degree offline decomposability judgements vary with age, as this might further affect the generalizability of many sets of idiom norms.

A comparison of the two studies (two-noun vs. one-noun idioms) revealed that the effects of Age and Frequency on the familiarity judgements in Studies 1 and 2 are roughly similar. The most important difference is a decrease in familiarity for the older participants (> 60 years) in Study 2, but not in Study 1. It is not clear what has caused this difference: The education levels of the older participants are very similar between studies (**Table 3**), and the predictors Education and Gender did not show an effect on the familiarity ratings in the statistical analyses. However, the number of older participants in Study 2 was much lower than in Study 1, and hence the variation between participants might have had a larger effect than in Study 1. In comparison, the difference in the effects of decomposability is relatively smaller.

Overall, the pattern of the idiom acquisition curve that we find in the two studies shows that—unlike what is often taken for granted in idiom processing studies—idiom knowledge varies widely between age groups. Especially young adolescents (students) cannot be expected to have developed a large idiom vocabulary yet. This implies that they constitute a relatively unreliable group for testing theories of idiom comprehension and production: they may or may not be familiar with the items, and their representations may be less stable than those of speakers above the age of 30.

TABLE 3 | Comparison of the levels of education of the older participants (> 60 years) in Study 1 and in Study 2.

	Study 1	Study 2
High school	6	4
Vocational education	9	4
University	21	11
Total	36	19

A possible explanation for the delay in idiom acquisition (in comparison to that of the single-word lexicon as described by Brysbaert et al., 2016) may be found both in the subject and item characteristics. First, idioms are figurative expressions and the ability to handle such expressions successfully only develops at around 9 years of age (Levorato and Cacciari, 1992). Second, idioms often tend to refer to relatively abstract and/or pragmatically complex concepts that may only be grasped well-beyond puberty. A third possibility is that what we observe in our data is in fact an indicator of language change. That is, the younger participants may simply not be familiar with the items because they are no longer being used and/or have been replaced by new idioms. Given the method of item selection (based on examples found in newspaper articles and conversations, but also idiom dictionaries), we do not find this explanation very likely, but we feel that it would be worth exploring in a future study. A methodological challenge will be however that any *new* idioms that would be expected to replace the old items in the vocabularies of the younger generations will first need to be identified.

At the other end of the distribution, old age, our findings are somewhat inconclusive: do elderly speakers experience problems in accessing items that they used to know before? Based on findings by Kuiper et al. (2009) and Escaip (2015) that showed mixed evidence for a late drop in idiom knowledge, we were especially interested in the category of 65+ participants. While our first study does not show any evidence for such a drop, the second study shows a slight decrease. Yet, the relatively few subjects in these categories and the large variability make it difficult to estimate the reliability of this effect. An additional explanation may be the influence of Frisian in this sample. Although we took care to remove all native speakers of Frisian, it is possible that the remaining participants are also predominantly located in Friesland and therefore come across different idioms in everyday life. For the older participants, this effect may be much stronger, as they can be expected to be less mobile and less exposed to mainstream (Dutch) media. In a future study, we therefore need to include information about the subject's geographical location, about the area in which they grew up, and the type of media that they consume. Ideally, this would be a megastudy comparable to that of Brysbaert et al. (2016), with a large number of items and a very large and diverse sample of Dutch speakers.

This study provides support for the hypothesis that idiom acquisition is similar to word acquisition, with increasing knowledge across the life span. However, idioms are different from words in that they are multiword expressions, and idioms are different from many other types of multiword expressions in that they have a figurative meaning. It would be interesting to compare the acquisition of idioms with the acquisition of other types of multiword expression to investigate how the ability to understand figurative expressions influences idiom acquisition. Is this a prerequisite for acquiring idioms, as is generally assumed? Or do children acquire high frequent idioms as words, without the ability to understand figurative language? One of the difficulties in investigating these questions is the variability between idioms. Other factors such as concreteness and imageability (both related to the transparency of the idiom) could play a role in whether and how much idioms are being perceived as figurative language. The effects that we find for decomposability support this hypothesis.

Taken together, our findings stress the need for future work to address both item and subject characteristics that could potentially affect idiom acquisition in more detail. Our findings with respect to the effect of decomposability and its interaction with age suggest that this could be a worthwhile enterprise. With respect to other item characteristics, possible candidates are, for example, the above mentioned factors concreteness, transparency, and imageability, but also length, or animacy.

While the idea that idioms differ with respect to item characteristics, such as decomposability, was formulated early on in the idiom literature (e.g., Gibbs and Nayak, 1989), the focus on speaker characteristics is relatively new (see also section Introduction). Yet, the idea that successful idiom comprehension depends on individual differences in processing abilities seems relatively straightforward, as idiom comprehension is a complex skill which is only acquired late during acquisition. For example, Cacciari et al. (2018) found a clear relationship between online idiom comprehension and cognitive functions that might extend to idiom acquisition as well, to the extent that it reflects differences in fluid intelligence. On top of that, the personality traits that they found to affect online processing might come into play in acquisition, too. The use of figurative language and other multi-word expressions is an important stylistic device that may be very well-suited to express different types of personality. The factor Age, which has been in the focus of the present article, may thus not only represent a participant's linguistic experience, but also its interaction with age-dependent changes in cognitive control, long-term memory access, and personality. Future studies will need to distinguish these factors on a more fine-grained level.

Are all native speakers alike when it comes to idioms? We have shown that—similar to the single-word vocabulary—the idiom vocabulary differs widely across speakers, with age rather than education being the main factor driving these differences. *Are all idioms alike when it comes to the probability of being known by a native speaker?* In line with findings on online processing (e.g., Arnon and Snider, 2010) we have shown that idioms behave much like ordinary entries in the mental lexicon, in that they are sensitive to distributional information. The more frequent an idiom, the larger the probability that a native speaker is familiar with it. In addition, the probability with which an idiom is acquired is affected by the degree to which it is decomposable. Our findings can help increase the reliability and validity of idiom processing studies. More importantly, we think that they contribute to a clearer picture of the way in which the boundaries of the lexicon expand across the lifespan.

DATA AVAILABILITY

The datasets generated for this study can be found here: https://git.lwp.rug.nl/p251653/development-idiom-knowledge.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the Ethical Rules for Conducting Research with Human Participants, Research Ethics Committee Faculty of Arts (CETO), University of Groningen. The protocol for all age groups was approved by the Research Ethics Committee Faculty of Arts (CETO), University of Groningen (60761519). All subjects gave written informed consent in accordance with the Declaration of Helsinki. We did not obtain written consent from parents/legal guardians of participants under 16, because participants participated anonymously in the experiment by clicking a link that was posted on social media, the participants were free to stop whenever they wished without consequences (they were not payed or reimbursed for their participation), and the materials of the experiment (officially recognized Dutch idioms) gave no reason to assume that non-adult participants could suffer negative consequences from participating in this study.

AUTHOR CONTRIBUTIONS

SS and JvR: conceptualization and writing. SS and AlR: materials and data collection. JvR: statistical analyses.

FUNDING

This research was supported by grants from the Netherlands Organization for Scientific Research NWO (Veni grant no. 275-70-044, JvR; and Ph.D. in the Humanities grant no. 322-75-008, AlR).

ACKNOWLEDGMENTS

We thank Prof. Dr. Gertjan van Noord and Peter Kleiweg for their assistance with the search for idioms in the Lassy Large corpus.

REFERENCES

Arnon, I., and Cohen Priva, U. (2013). More than words: the effect of multi-word frequency and constituency on phonetic duration. *Lang. Speech* 56, 349–371. doi: 10.1177/0023830913484891

Arnon, I., and Snider, N. (2010). More than words: frequency effects for multi-word phrases. *J. Memory Lang.* 62, 67–82. doi: 10.1016/j.jml.2009.09.005

Bannard, C., and Matthews, D. (2008). Stored word sequences in language learning: the effect of familiarity on children's repetition of four-word combinations. *Psychol. Sci.* 19, 241–248. doi: 10.1111/j.1467-9280.2008.02075.x

Bobrow, S. A., and Bell, S. M. (1973). On catching on to idiomatic expressions. *Mem. Cogn.* 1, 343–346.

Bonin, P., Méot, A., Boucheix, J. M., and Bugaiska, A. (2017). Psycholinguistic norms for 320 fixed expressions (idioms and proverbs) in French. *Quart. J. Exp. Psychol.* 71:1057–69. doi: 10.1080/17470218.2017.1310269

Bonin, P., Méot, A., and Bugaiska, A. (2013). Norms and comprehension times for 305 French idiomatic expressions. *Behav. Res. Methods* 45, 1259–1271. doi: 10.3758/s13428-013-0331-4

Brysbaert, M., Stevens, M., Mandera, P., and Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Front. Psychol.* 7:1116. doi: 10.3389/fpsyg.2016.01116

Bulkes, N. Z., and Tanner, D. (2017). "Going to town": large-scale norming and statistical analysis of 870 American English idioms. *Behav. Res. Methods* 49, 772–783. doi: 10.3758/s13428-016-0747-8

Cacciari, C., Corradini, P., and Ferlazzo, F. (2018). Cognitive and personality components underlying spoken idiom comprehension in context. An exploratory study. *Front. Psychol.* 9:659. doi: 10.3389/fpsyg.2018.00659

Cacciari, C., Padovani, R., and Corradini, P. (2007). Exploring the relationship between individuals' speed of processing and their comprehension of spoken idioms. *Eur. J. Cogn. Psychol.* 19, 417–445. doi: 10.1080/09541440600763705

Cacciari, C., and Tabossi, P. (1988). The comprehension of idioms. *J Memory Lang.* 27, 668–683. doi: 10.1016/0749-596X(88)90014-9

Caillies, S. (2009). Descriptions de 300 expressions idiomatiques: familiarité, connaissance de leur signification, plausibilité littérale,≪ décomposabilité ≫ et ≪ prédictibilité ≫. *L'année Psychol.* 109, 463–508. doi: 10.4074/S0003503309003054

Caillies, S., and Butcher, K. (2007). Processing of idiomatic expressions: evidence for a new hybrid view. *Metaphor. Symbol.* 22, 79–108. doi: 10.1080/10926480709336754

Cain, K., Oakhill, J., and Lemmon, K. (2005). The relation between children's reading comprehension level and their comprehension of idioms. *J. Exp. Child Psychol.* 90, 65–87. doi: 10.1016/j.jecp.2004.09.003

Citron, F. M., Cacciari, C., Kucharski, M., Beck, L., Conrad, M., and Jacobs, A. M. (2016). When emotions are expressed figuratively: psycholinguistic and Affective Norms of 619 Idioms for German (PANIG). *Behav. Res. Methods* 48, 91–111. doi: 10.3758/s13428-015-0581-4

Columbus, G., Sheikh, N. A., Côté-Lecaldare, M., Häuser, K., Baum, S. R., and Titone, D. (2015). Individual differences in executive control relate to metaphor processing: an eye movement study of sentence reading. *Front. Hum. Neurosci.* 8:1057. doi: 10.3389/fnhum.2014.01057

Cutting, J. C., and Bock, K. (1997). That's the way the cookie bounces: syntactic and semantic components of experimentally elicited idiom blendsß. *Memory Cogn.* 25, 57–71. doi: 10.3758/BF03197285

Escaip, V. (2015). *The Relationship of Phrase Head Word Frequency and Acquired Idioms: A Comparative Analysis of Spanish, English and French Verb Phrase Idioms*. Unpublished thesis, University of Canterbury, New Zealand. Retrieved from: http://hdl.handle.net/10092/11757

Fillmore, C. J., Kay, P., and O'connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: the case of let alone. *Language* 64, 501–538.

Gibbs, R. W. (1991). Semantic analyzability in children's understanding of idioms. *J. Speech Language Hearing Res.* 34, 613–620. doi: 10.1044/jshr.3403.613

Gibbs, R. W. Jr., and Nayak, N. P. (1989). Psycholinguistic studies on the syntactic behavior of idioms. *Cogn. Psychol.* 21, 100–138. doi: 10.1016/0010-0285(89)90004-2

Gibbs, R. W. Jr., Nayak, N. P., and Cutting, C. (1989). How to kick the bucket and not decompose: analyzability and idiom processing. *J. Memory Lang.* 28, 576–593. doi: 10.1016/0749-596X(89)90014-4

Hastie, T., and Tibshirani, R. (1990). *Generalized Additive Models. Monographs on Statistics and Applied Probability,* Vol. 43. London: Chapman and Hall.

Hung, P. F., and Nippold, M. A. (2014). Idiom understanding in adulthood: Examining age-related differences. *Clin. Lingu. Phonet.* 28, 208–221. doi: 10.3109/02699206.2013.850117

Jackendoff, R. (1995). "The boundaries of the lexicon," in *Idioms: Structural and Psychological Perspectives,* eds M. Everaert, E. J. Van der Linden, R. Schreuder, and R. Schreuder (Hillsdale, NJ: Lawrence Erlbaum Associates), 133–166.

Jackendoff, R. (1997). Twistin'the night away. *Language* 73, 534–559.

Janssen, N., and Barber, H. A. (2012). Phrase frequency effects in language production. *PLoS ONE* 7:e33202. doi: 10.1371/journal.pone.0033202

Kuiper, K., Columbus, G., and Schmitt, N. (2009). "The acquisition of phrasal vocabulary," in *Language Acquisition*, eds Foster-Cohen and Susan (London: Palgrave Macmillan), 216–240. doi: 10.1057/9780230240780_10

Levorato, M. C., and Cacciari, C. (1992). Children's comprehension and production of idioms: the role of context and familiarity. *J. Child Lang.* 19, 415–433. doi: 10.1017/S0305000900011478

Li, D., Zhang, Y., and Wang, X. (2016). Descriptive norms for 350 Chinese idioms with seven syntactic structures. *Behav. Res. Methods* 48, 1678–1693. doi: 10.3758/s13428-015-0692-y

Libben, M. R., and Titone, D. A. (2008). The multidetermined nature of idiom processing. *Mem. Cogn.* 36, 1103–1121. doi: 10.3758/MC.36.6.1103

Moon, R. (1998). *Fixed Expressions and Idioms in English : A Corpus-Based Approach. Oxford Studies in Lexicography and Lexicology.* Oxford: Clarendon.

Nippold, M. A., and Duthie, J. K. (2003). Mental imagery and idiom comprehension: a comparison of school-age children and adults. *J. Speech Lang. Hear. Res.* 46, 788–799. doi: 10.1044/1092-4388(2003/062)

Nippold, M. A., and Martin, S. T. (1989). Idiom interpretation in isolation versus context: a developmental study with adolescents. *J. Speech Lang. Hear. Res.* 32, 59–66. doi: 10.1044/jshr.3201.59

Nippold, M. A., and Rudzinski, M. (1993). Familiarity and transparency in idiom explanation: a developmental study of children and adolescents. *J. Speech Lang. Hear. Res.* 36, 728–737. doi: 10.1044/jshr.3604.728

Nippold, M. A., and Taylor, C. L. (1995). Idiom understanding in youth: Further examination of familiarity and transparency. *J. Speech Lang. Hear. Res.* 38, 426–433. doi: 10.1044/jshr.3802.426

Nordmann, E., and Jambazova, A. A. (2017). Normative data for idiomatic expressions. *Behav. Res. Methods* 49, 198–215. doi: 10.3758/s13428-016-0705-5

Pawley, A., and Syder, F. H. (1983). "Two puzzles for linguistic theory: Nativelike selection and nativelike fluency," in *Language and Communication*, eds J. C. Richards and R. W. Schmidt (New York, NY: Longman Inc,), 191–226.

R Core Team (2018). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna. Available online at: https://www.R-project.org/

Rommers, J., Dijkstra, T., and Bastiaansen, M. (2013). Context-dependent semantic processing in the human brain: evidence from idiom comprehension. *J. Cogn. Neurosci.* 25, 762–776. doi: 10.1162/jocn_a_00337

Sprenger, S., and van Rijn, H. (2013). It's time to do the math: Computation and retrieval in phrase production. *Mental Lexicon* 8, 1–25. doi: 10.1075/ml.8.1.01spr

Sprenger, S. A., Levelt, W. J., and Kempen, G. (2006). Lexical access during the production of idiomatic phrases. *J. Memory Lang.* 54, 161–184. doi: 10.1016/j.jml.2005.11.001

Swinney, D. A., and Cutler, A. (1979). The access and processing of idiomatic expressions. *J. Verbal Learn. Verbal Behav.* 18, 523–534. doi: 10.1016/S0022-5371(79)90284-6

Tabossi, P., Arduino, L., and Fanari, R. (2011). Descriptive norms for 245 Italian idiomatic expressions. *Behav. Res. Methods* 43, 110–123. doi: 10.3758/s13428-010-0018-z

Titone, D., and Libben, M. (2014). Time-dependent effects of decomposability, familiarity and literal plausibility on idiom priming: a cross-modal priming investigation. *Mental Lexicon* 9, 473–496. doi: 10.1075/ml.9.3.05tit

Titone, D., Lovseth, K., Kasparian, K., and Tiv, M. (2019). Are figurative interpretations of idioms directly retrieved, compositionally built, or both? Evidence from eye movement measures of reading. *Can. J. Exp. Psychol.* doi: 10.1037/cep0000175

Titone, D. A., and Connine, C. M. (1994). Descriptive norms for 171 idiomatic expressions: familiarity, compositionality, predictability, and literality. *Metaphor. Symbol.* 9, 247–270. doi: 10.1207/s15327868ms0904_1

Titone, D. A., and Connine, C. M. (1999). On the compositional and noncompositional nature of idiomatic expressions. *J. Prag.* 31, 1655–1674.

Tremblay, A., and Tucker, B. V. (2011). The effects of N-gram probabilistic measures on the recognition and production of four-word sequences. *Mental Lexicon* 6, 302–324. doi: 10.1075/ml.6.2.04tre

Van Noord, G., Bouma, G., Van Eynde, F., De Kok, D., Van der Linde, J., Schuurman, I., et al. (2013). "Large scale syntactic annotation of written Dutch: Lassy," in: *Essential Speech and Language Technology for Dutch* (Berlin; Heidelberg: Springer), 147–164. Available online at: https://www.let.rug.nl/vannoord/Lassy/

van Rij, J., Hendriks, P., van Rijn, H., Baayen, R. H., and Wood, S. N. (2019). Analyzing the time course of pupillometric data. *Trends Hear. Sci.* 23:2331216519832483. doi: 10.1177/2331216519832483

van Rij, J., Vaci, N., Wurm, L. H., and Feldman, L., B. (in press) "Alternative quantitative methods in psycholinguistics: Implications for theory and design," in *Word Knowledge and Word Usage: a Cross-disciplinary Guide to the Mental Lexicon,* eds V. Pirrelli, I. Plag, and W. U. Dressler (Berlin: Mouton de Gruyter).

van Rij, J., Wieling, M., Baayen, R. H., and van Rijn, H. (2017). *itsadug: Interpreting Time Series and Autocorrelated Data Using Gamms*. Published on the Comprehensive R Archive Network (CRAN). Available online at: https://cran.r-project.org/web/packages/itsadug.

Westbury, C., and Titone, D. (2011). Idiom literality judgments in younger and older adults: age-related effects in resolving semantic interference. *Psychol. Aging* 26, 467–474. doi: 10.1037/a0022438

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: a tutorial focusing on articulatory differences between l1 and l2 speakers of english. *J. Phonet.* 70, 86–116. doi: 10.1016/j.wocn.2018.03.002

Wood, S. (2017). *Generalized Additive Models: An Introduction With R, Second Edition. Chapman & Hall/CRC Texts in Statistical Science*. Boca Raton, FL: CRC Press.

Wood, S. N., Pya, N., and Saefken, B. (2016). Smoothing parameter and model selection for general smooth models. *J. Am. Statist. Assoc.* 111, 1548–1575. doi: 10.1080/01621459.2016.1180986

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 09 July 2019
doi: 10.3389/fpsyg.2019.01570

Check for updates

If Birds Have Sesamoid Bones, Do Blackbirds Have Sesamoid Bones? The Modification Effect With Known Compound Words

Thomas L. Spalding[1], Christina L. Gagné[1], Kelly A. Nisbet[2], Jenna M. Chamberlain[1] and Gary Libben[3]*

[1] Department of Psychology, University of Alberta, Edmonton, AB, Canada, [2] Department of Linguistics and Languages, McMaster University, Hamilton, ON, Canada, [3] Department of Psychology, Brock University, St. Catharines, ON, Canada

Three experiments investigate how people infer properties of compound words from the unmodified head. Concepts license inference of properties true of the concept to instances or sub-types of that concept: Knowing that birds generally fly, one infers that a new type of bird flies. However, different names are also believed to reflect real underlying differences. Hence, a different name creates the expectation that a new bird differs from birds in general, and this might impact property inference. In these experiments, participants were told, Almost all (Some, Almost no) birds have sesamoid bones, and then asked, What percentage of blackbirds (birds) have sesamoid bones? The results indicate both inference and contrast effects. People infer properties as less common of the compound than the head when the property is true of the head, but they infer them as more common of the compound than the head when the property is not true of the head. In addition, inferences about properties true of the head are affected by the semantic similarity between the head and the compound, but properties not true of the head do not show any semantic similarity effect, but only a small, consistent effect of contrast. Finally, the presentation format (Open vs. Closed compounds) affects the pattern of effects only when the spacing suggests the existence of a permanent name.

Keywords: modification effect, compound words, modifier-noun phrases, property verification, concepts

OPEN ACCESS

Edited by:
Guillaume Thierry, Bangor University, United Kingdom

Reviewed by:
Giorgio Arcara, IRCCS Fondazione Ospedale San Camillo, Italy
Ernesto Guerra, Universidad de Chile, Chile

****Correspondence:***
Thomas L. Spalding
spalding@ualberta.ca

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *29 October 2018*
Accepted: *21 June 2019*
Published: *09 July 2019*

Citation:
Spalding TL, Gagné CL, Nisbet KA, Chamberlain JM and Libben G (2019) If Birds Have Sesamoid Bones, Do Blackbirds Have Sesamoid Bones? The Modification Effect With Known Compound Words. Front. Psychol. 10:1570. doi: 10.3389/fpsyg.2019.01570

INTRODUCTION

Much research on compound words (words that consist of two or more free morphemes, e.g., *snowball* or *hogwash*) focuses on the processing involved in accessing the words (see, e.g., Libben, 1998; Libben and Jarema, 2006, for reviews) for use in specifically language-related tasks. There is much less work on how compound words are used more broadly in human cognition. In this paper, we investigate how compound words function in, and contribute to, human cognition more broadly. In particular, we are interested in what support compound words provide to conceptual (or categorical) inference. It is well accepted that a major function of concepts in human thinking is to provide the ability to infer properties from the concept to members of the category named or referred to by the concept (see e.g., Murphy, 2002; see also, Osherson et al., 1990). Thus, if one knows that birds have sesamoid bones, then one can make a reasonable inference that a particular new bird has sesamoid bones, even though one has no other information that would specifically indicate this fact for this particular bird. Similarly, one might make a reasonable inference that blackbirds, in general, have sesamoid bones, if birds, in general, are believed to have sesamoid bones. Importantly, such inferences are probabilistic and defeasible, rather than strictly logical deductions (see Osherson et al., 1990), such that, for example, specific information about a particular item or sub-group may override the inference (e.g., knowledge that penguins cannot fly rules out the inference that penguins can fly, just because they are a kind of bird), as can the typicality of the sub-group with respect to the group (e.g., property inference is more likely from bird to robin than from bird to, say, turkey, even in the absence of knowledge about the specific property). However, the point of categorical inference is precisely that the category allows one to infer properties where there is no specific knowledge about the item or sub-group that can be brought to the question.

On the other hand, it is equally well accepted that people expect that different names for things reflect real underlying semantic differences. Indeed, the literature contains three proposed principles of human cognition (Synonomy Avoidance, Carstairs-McCarthy, 2010; Principle of Contrast, Clark, 1993; and Mutual Exclusivity Principle, Markman, 1989), each of which shares this core notion of expectation that different names reflect real underlying differences. While these principles are generally framed in terms of when a new name is, in some sense, justified, and how people make that decision, it is also the case that when a new name is presented as established, this leads to an expectation of some real differences from things that already have other names. Importantly, this expectation seems to be quite general, arising even when there is no specific knowledge of an existing difference. Taking these principles into account, then, one might expect that birds and blackbirds should be assumed to differ in significant ways, and this expectation might then affect the process of inferring properties from the head to the compound word. Thus, perhaps although birds generally do have sesamoid bones, blackbirds might not.

Clearly, then, we have two well-attested principles, which seem to work in opposite directions in terms of property inference: That categories license inference of properties to new sub-categories or category members, and that new labels, for example, of a sub-category, indicate property differences (i.e., a lack of licensed inference) for new sub-categories or category members. It is important to understand how these two principles operate together, as property inference is a major communicative function: Property inference (or lack thereof) creates expectations of novel objects, people, situations, and so on, based on what is known or said of an existing set, so that one is in a better position to deal with that novel object, person, or situation. This predictive function is also critical in numerous areas of applied work, for example in natural language processing, where understanding the expectations of a user is critical to the success of the system. As one example, research aimed at improving information retrieval engines (e.g., Baldwin et al., 2010; Wang et al., 2010) by identifying possible alternative terms that might be used to facilitate information access is one example where it is critical to understand what humans expect about the meaning of new terms. Similarly, understanding property inference can contribute to research on natural language question answering (for an overview, see Hirschman and Gaizauskas, 2001). For example, Wang et al. (2011) presented methods for identifying discourse structure for online forum data (see also Wang et al., 2010). The dialogue acts within these threads have various structures such as Question–Question, Answer–Answer, Question–Additional Information, and the meaning of terms in these dialog acts depends, in part, on the structure of those dialog acts (e.g., similar expressions can carry somewhat different meanings and referents, depending on whether they are embedded in an answer or in a request for additional information). Discovering the factors that influence the way in which the expression is used to refer to a referent and the properties of that referent can aid the development of NLP procedures used to automate the identification of dialogue acts. In addition, systems that attempt to automatically extract sentiment must be built keeping in mind the ways in which people infer properties of objects as existing labels are used and new labels are introduced (e.g., Maynard and Funk, 2012; Dragos et al., 2018). In sum, understanding how property inference operates in light of these two contrasting principles has important scientific, and also applied, consequences.

We begin by reviewing what is known about how people extend properties of concepts to novel combined concepts, and then turn to the question of whether people extend properties to known compound words in the same way. The question of when and which properties of a combined concept become available during conceptual combination has been one of the core questions within the conceptual combination literature. Research on this topic initially examined whether properties of the constituent concepts are available prior to properties of the whole concept. Early research by Springer and Murphy (1992) found that people were faster to verify properties that were true of the phrase (e.g., peeled apples are white) than properties that were true of the head concept, prior to modification (e.g., peeled apples are round). Gagné and Murphy (1996) found that discourse context did not alter this pattern.

More recently, the question of property inclusion has been examined in the context of examining whether the availability of properties differs for the head noun concept (e.g., ducks have webbed feet) relative to a modified concept (e.g., baby ducks have webbed feet). When using novel combinations, this work reveals a robust set of effects called the modification and inverse modification effects (see Spalding and Gagné, 2015, for a demonstration of both modification and inverse modification effects, but for modification effects using other property verification tasks see also Connolly et al., 2007; Jönsson and Hampton, 2008, 2012; Gagné and Spalding, 2011, 2014b; Hampton et al., 2011). In particular, properties generally true of the unmodified head noun become less true of the modified head (modification effect), while properties generally false of the unmodified head become less false of the modified head (inverse modification effect). Thus, for example, purple candles are judged less likely to be made of wax than candles, but purple candles are also judged more likely to have teeth than candles. The modification effect is robust over a range of specific verification tasks, including ratings of likelihood of the truth/plausibility of a property for a category (e.g., Connolly et al., 2007), true/false decisions about the property's relation to the concept and the response times to make those decisions (e.g., Gagné and Spalding, 2011), and estimates of the percentages of category members for which the property is true (e.g., Spalding and Gagné, 2015). Also, the modification effect is very robust over a wide range of property typicality, including properties that seem to be nearly definitional of the head, such as being animate for lamb (see e.g., Jönsson and Hampton, 2008, 2012; Hampton et al., 2011). This robustness over various kinds of properties is unexpected by those theories where prototypicality of features should be a determining aspect of property verification, such as prototype theories of conceptual combination (e.g., Hampton, 1991), but also theories of the semantics of compound words which differentiate between "levels" of properties (e.g., the skeleton vs. body distinction in Lieber, 2004).

Although the modification effect was initially used to test hypotheses about whether or not properties of a combined concept are directly inherited from the constituent concepts (e.g., Connolly et al., 2007), there are some findings to suggest that this effect might not actually be driven by the process of conceptual combination (i.e., constructing a new concept based on the conceptual "contents" of the modifier and head) *per se*, but rather by reasoning about the combined concepts. Gagné and Spalding (2011, 2014b, 2015), Spalding and Gagné (2015), Gagné et al. (2017) present evidence that the modification effects primarily arise due to meta-knowledge of modification, and particularly to reasoning about the expected relation of the meaning of the combined concept and the head. In particular, they argue that the effects are largely driven by the expected level of contrast (i.e., matching or mismatching features) between the combined concept and the head, rather than by conceptual knowledge of the individual constituent concepts (as would be expected by e.g., Hampton, 1987, 1991; but also by many approaches to the semantics of known compound words such as Lieber, 2004, 2009; see Gagné and Spalding, 2015, for a discussion). For example, modification effects arise even when the modifier is a non-word, and thus cannot contribute any semantic or conceptual information about what properties are appropriate for the combined concept (e.g., Spalding and Gagné, 2015).

In short, the literature on the modification effect strongly suggests that (a) the inferential function of concepts does indeed extend to modified versions of those concepts, rather than only to individual members of the category picked out by the concept, but that (b) the modification and inverse modification effects are driven by people's expectations about the nature, purpose, and use of modification. Thus, it seems likely that the general principle that a different name implies other, underlying differences (Synonomy Avoidance, Carstairs-McCarthy, 2010; Principle of Contrast, Clark, 1993; and Mutual Exclusivity Principle, Markman, 1989) should lead to modification and inverse modification effects with known compounds. However, in the existing literature, the "different names" created by modification are not well-established, but rather are novel. This novelty could have two different kinds of influence on the modification effect. It could be that novel names are simply seen as less established or less permanent names, and therefore they might lead to smaller modification effects (i.e., more likelihood of property inference), or it could be that the novelty makes the contrast more immediately salient, and therefore leads to larger modification effects (i.e., less likelihood of property inference). Furthermore, because of the novelty of the modified concepts, the existing literature is unable to investigate the way in which these principles interact with the semantic knowledge that is inherent in a category of things with a well-established name, or, indeed, with the simple fact of the well-established name. Nevertheless, there is good reason to believe that much of the processing and semantics of compound words is similar to that of novel conceptual combinations (see, e.g., Gagné and Spalding, 2014a). Thus, we expect to see modification effects (and inverse modification effects) with known compounds.

In the current experiments, we investigate the seeming conflict between the basic cognitive principles that we infer properties based on category membership and that the use of different names for things implies real underlying differences (Synonomy Avoidance, Carstairs-McCarthy, 2010; Principle of Contrast, Clark, 1993; and Mutual Exclusivity Principle, Markman, 1989), in three experiments. In particular, we investigate the extent to which people are willing to infer properties from head nouns to compounds under various conditions that should affect the extent to which people believe the compounds to be well-established as different names, and thus should bring the principle that different names imply different properties more into conflict with the principle that properties can be inferred from categories to more specific sub-sets. Experiment 1 investigates whether the modification and inverse modification effects occur for known, transparent compound words and further investigates whether semantic knowledge of the compounds and their relation to the head nouns plays a role. Experiment 2 replicates Experiment 1, and in addition investigates the effect of spacing, under the assumption that the lack of spacing would indicate the compounds as more established, permanent, unique names for existing sub-categories. Experiment 3 is similar to Experiment 2, except that we replaced the modifier of the compound with a non-word, in order to find out if the structural cue (spacing) would affect property inference when the "compound" is not a known category.

EXPERIMENT 1

This experiment investigates how people infer properties of transparent compound words based on their knowledge of those properties' relationship with the head noun, using the method from Spalding and Gagné (2015). This method asks participants to estimate the percentage of members of a category (e.g., buds) or subcategory (e.g., rosebuds) that have a particular property.

To directly manipulate the truth value of the property, we used blank predicates (i.e., properties that use relatively unfamiliar terms but are relevant for the concept in question, e.g., biological predicates are used for animal categories, see

Osherson et al., 1990). The likelihood of the property for the unmodified noun was manipulated by telling participants that Almost All, Some, or Almost No members of the head noun concept had it. Critically, because we used the same blank properties in the Almost all, Some, and Almost No conditions, not only is the semantic content of those properties relatively unfamiliar and/or unrelated to the compounds, but to the extent it is familiar, the semantic content is controlled across the likelihood manipulation.

Based on previous work showing that processing compound words appears to involve many of the same processes as the processing of novel combined concepts (e.g., Gagné and Spalding, 2009), we predict that in making a decision about whether a property of the head is true of a compound word, people will show a pattern similar to that previously demonstrated for novel modifier-noun pairings: Modification effect for properties true of the head and inverse modification effect for properties false of the head.

Methods

This and the following experiments were reviewed for ethical content and approved by the Research Ethics Committee at the University of Alberta, and all participants provided written consent.

Participants

Sixty-two participants took part in the study. Participants in this and all following experiments were undergraduate students enrolled in a very large first year psychology class and obtained partial class credit for participating. In this participant population, approximately 95% are between 18 and 24 years of age and approximately 70% are female. All participants in this and the following experiments self-identify as native English speakers. In this, and the following experiments, target sample size is determined by expected effect size and complexity of design. The number of actual participants is determined by the target sample size and availability of participants in the Departmental participant pool. In this experiment, we set a target of 10 participants per condition. Two extra participants were included (in our pool, if extra participants attend a session, they must be run).

Materials and Design

Experiment 1 used 96 transparent compounds (e.g., *snakeskin*) selected from a previously categorized set of items (Ji et al., 2011). Statements to be predicated of the heads of the compounds were then selected to match each compound word. The truth of these statements, relative to the compound words or the head nouns of the compound words, were expected to be unknown by the participants. The unknown predicates were taken from previous experiments on the modification effect (Spalding and Gagné, 2015) and from Wikipedia searches for uncommon words related to the head noun. The compounds and the predicates are presented in the **Appendix**. The design is a 2 (Modification: modified vs. unmodified) by 3 (Likelihood: almost all, some, almost no) crossed factorial design.

Procedure

On each trial, participants were first shown a statement regarding how often an unknown property is true of an unmodified noun using one of three quantifiers: Almost all, Some, or Almost no. For example, participants might see "Almost all birds require graminoids in their diet." The participants were instructed to treat this statement as true. Participants were then asked a follow up question about how many members of the unmodified noun or the modified noun category have that same property. They were asked to respond on a scale of 1–100. For example, they would be asked either "What percentage of birds require graminoids in their diet?" or "What percentage of blackbirds require graminoids in their diet?". The lists were counterbalanced so that each participant saw either the compound word or the head noun in the question.

Results and Discussion

Data Analysis

The descriptive statistics are shown in **Table 1**. We analyzed the data using linear-mixed effects (LME) regression models in which subject and item were entered as random effects, and Modification (modified vs. unmodified) and Likelihood (Almost all, Some, Almost no) were entered as fixed effects, using the *mixed* and *contrast* commands in Stata (StataCorp, 2017). The *mixed* function outputs coefficients (i.e., estimates) for simple effects at the first level of the other categorical variables and at the mean of the other continuous variables in the model (see **Table 2**). For testing our hypotheses, these coefficients are not directly interpretable because they represent the simple effect of a variable at the first level of other variables. The relevant statistical tests for addressing our research questions concern interactions and simple effects, which are reported in the following text. The *contrast* function in Stata was used to conduct these analyses. We report the tests conducted on these fixed effects. Tests of simple effects (the effect of a factor at one level of another factor) were conducted to follow up on statistically significant interactions, because in the case of significant interactions, the main effects are not informative. Some statistics packages report LME main effects as *F*-tests and the simple effects as *t*-tests. However, because the degrees of freedom are indeterminate for such tests in linear mixed effect models, Stata uses chi-square and *Z*-scores, respectively.

Results

The analysis indicated a significant interaction between Modification and Likelihood, $X^2(2) = 435.0$, $p < 0.001$. Analysis of the simple effects indicated that the Almost All and Some

TABLE 1 | Mean (SE) judged percentage of category members having the test property by level of Likelihood from Experiment 1.

Condition	Percent of category members (SE)		
	Almost All	Some	Almost No
Unmodified	91.1 (1.5)	37.6 (2.0)	7.2 (1.4)
Modified	66.7 (4.7)	28.2 (3.0)	9.8 (2.3)

TABLE 2 | Experiment 1 Mixed Model Coefficients.

Variable	Coefficient	Standard Error	z	P > |z|
Likelihood: Almost no	−56.8	0.916	−62.0	0.000
Likelihood: Some	−38.6	0.918	−42.1	0.000
Modification	24.4	0.909	26.9	0.000
Mod × Likelihood: Almost no	−27.0	1.29	−20.8	0.000
Mod × Likelihood: Some	−15.0	1.30	−11.6	0.000
Constant	66.7	1.03	65.0	0.000
Subjects	31.8	6.44		
Items	13.0	2.85		

conditions both led to a significant modification effect, $z = 26.9$, $p < 0.001$ and $z = 10.1$, $p < 0.001$, respectively, while the Almost No condition led to a significant inverse modification effect, $z = -2.8$, $p < 0.01$. That is, the property was judged less likely for the compound, if the property was presented as true of almost all or some of the head category members, but more likely for the compound, if the property was presented as true of almost no members of the head category. Clearly, known compound words give rise to a very robust modification and inverse modification effect, as predicted.

These results show that modification and inverse modification effects that have been reported for novel phrases extend to lexicalized compounds. However, there are several more specific points that should be noted. First, the modification effects are much larger than those found in previous studies with novel phrases. For example, in the current experiment, modification in the Almost All condition reduced the likelihood of the property by some 24 percentage points, while Spalding and Gagné (2015) found a reduction of only 2 percentage points, using exactly the same experimental paradigm and participant population. This is consistent with the notion that the degree to which the compound word is seen as a unique, permanent, established name is likely to make people believe more strongly that there are real, underlying differences between the things named by the head and by the compound. Second, there is a strong asymmetry between the true and false features (i.e., between the size of the modification and inverse modification effects), suggesting that the existence of a known compound affects the features presented as generally false of the head noun much less than those presented as generally true of the head.

To investigate the pattern of asymmetry between the true and false features further, and to investigate the role of the semantics of the existing compound word, we performed a *post hoc* analysis, in which we entered the semantic similarity between the known compound and the unmodified head, as measured by Latent Semantic Analysis (LSA; Landauer and Dumais, 1997), into an LME regression with the Likelihood factor (Almost All, Some, Almost No), using only the percentage estimates for the Modified condition (see **Table 3**). Given that the compound was never seen in the unmodified condition, it would not be meaningful to include these items when examining the impact of the similarity between the compound and head. That is, the unmodified condition is when the head is presented for the property judgment, thus those data points reflect the judgment made when the participant is not presented with the compound word, and hence the semantic similarity between the compound and head is not meaningful in these cases (e.g., if the participant judges the likelihood that birds require graminoids in their diet, they will not have seen blackbird in the experiment at all, and hence the relationship between blackbird and bird is completely irrelevant). We observed a significant interaction between the LSA measure (e.g., the association between the word *birds* and the word *blackbirds*) and Likelihood, $\chi^2(2) = 38.2$, $p < 0.001$. Further analysis indicated a significant slope for LSA in the Almost All and Some conditions, $z = 5.8$, $p < 0.001$ and $z = 3.0$, $p < 0.005$, respectively. However, the semantic similarity between the compound and the unmodified head did not affect the estimates in the Almost No condition, $z < 1$. See **Figure 1**. In short, the modification effects were significantly smaller for compounds that were more semantically similar to their heads in the Almost All and Some conditions, but semantic similarity between the compound and the unmodified head had no effect in the Almost No condition.

To further examine whether property inference is influenced by degree of semantic transparency, we also used semantic transparency ratings by human participants from a database containing semantic transparency ratings for over 8000 compounds (Gagné et al., in press). The judgment that is relevant for the current experiment is the rating between the head noun and the compound. Participants were asked to judge on a scale from 0 to 100% how much the head noun retained its meaning in the compound (e.g., *How much is the meaning of birds retained in the meaning of blackbirds*). This information was available for 71 items. The mean rating was 79% ($SD = 13$) and ranged from 52 to 96%. This judgment was entered into a model that also included Likelihood and was restricted to only the modified concept condition (see **Table 4**). There was an interaction between transparency rating and Likelihood, $\chi^2(2) = 49.99$, $p < 0.0001$. Analysis of this interaction indicated a

TABLE 3 | Experiment 1 Mixed Model Coefficients with LSA: Modified condition only.

Variable	Coefficient	Standard Error	z	P > |z|
Likelihood: Almost no	−49.2	6.68	−24.8	0.000
Likelihood: Some	−34.2	1.98	−17.4	0.000
LSA	38.9	6.68	5.8	0.000
LSA × Likelihood: Almost no	−41.6	6.74	−6.2	0.000
LSA × Likelihood: Some	−18.8	6.94	−2.7	0.007
Constant	59.1	2.16	27.4	0.000
Subjects	2.1	0.12		
Items	2.0	0.12		

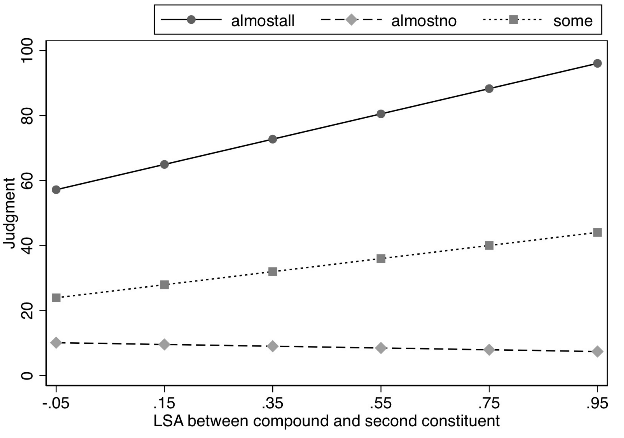


FIGURE 1 | Effect of semantic similarity (LSA) on judged likelihood of property in Experiment 1.

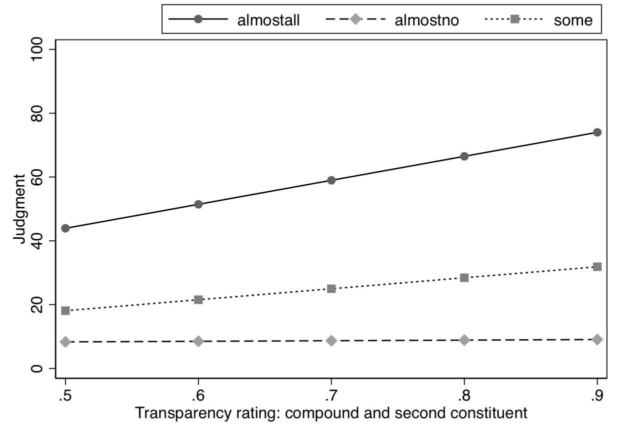


FIGURE 2 | Effect of semantic similarity (human ratings) on judged likelihood of property in Experiment 1.

significant slope for the transparency rating in the Almost All and Some conditions, $z = 7.77$, $p < 0.0001$ and $z = 3.32$, $p < 0.002$, respectively, with the slope for transparency judgments being steeper in the Almost All condition than in the Some condition, $\chi^2(1) = 13.83$, $p < 0.001$. As transparency increased, ratings for the properties increased for the modified items, meaning that more transparency would correspond to smaller modification effects. However, transparency judgments did not affect the estimates in the Almost No condition, $z = 0.19$, $p = 0.852$. See **Figure 2**.

Discussion

Although the specific properties being tested are unrelated to participants' existing knowledge of the known compound, properties presented as true or somewhat true of the head are sensitive to the overall known semantic difference between the compound and the head. Thus, Experiment 1 shows clearly that, in addition to an expectation of some contrast due to the existence of the compound as a name that contrasts in some way with the head, the semantics of the specific compound word also contribute to the modification effects with known compound words when Almost All or Some members of the head noun have that property.

However, when the properties are presented as false of the head (i.e., Almost No members of the head noun have the property) the participants are insensitive to the degree of semantic difference between the compound and the head, indicating that the inverse modification effect with such features might provide us with a kind of baseline measure of the pure effect of expectation of contrast driven purely by the fact of the different name (i.e., the fact that there is an existing compound that contrasts with the head). We propose that this difference stems from meta-knowledge about the relationship between properties and concepts. True properties are assumed to be related to the semantics of the concept and false properties are not (see e.g., Murphy and Medin, 1985, for discussion of what features are relevant to a given concept). That is, our concepts, in general, tend to be organized around things that are true of them (e.g., we tend to think of birds, for example, more in terms of the fact that they usually fly, have wings, and have feathers, rather than in terms of the fact that they do not usually explode or earn PhD's).

Thus, true properties are influenced by the actual (pre-existing) semantic similarity between the compound and head because this similarity is used as one source of information about the expected level of contrast when making property judgments in the Almost All and Some conditions. On the other hand, in the Almost No condition, semantic similarity is not seen as being relevant due to the meta-knowledge that false properties are not generally associated with the concepts. We return to this point in the section "General Discussion."

TABLE 4 | Experiment 1 Mixed Model Coefficients with Semantic Transparency judgments: Modified condition only.

Variable	Coefficient	Standard Error	z	P > |z|
Likelihood: Almost no	1.1	8.3	0.1	0.900
Likelihood: Some	−5.4	8.7	−0.6	0.534
Semantic Transparency	75.2	9.7	7.8	0.000
ST × Likelihood: Almost no	−73.4	10.4	−7.1	0.000
ST × Likelihood: Some	−40.8	11.0	−43.7	0.000
Constant	6.3	7.8	0.81	0.416
Subjects	2.1	0.11		
Items	1.9	0.13		

EXPERIMENT 2

Experiment 1 showed very robust modification and inverse modification effects, with a strong asymmetry between the size of the two effects. In this experiment, we attempt to replicate these effects. Further, we examine whether property inferences for known compounds are affected by presenting the compounds with an open structure (e.g., *black bird*) or a closed structure (e.g., *blackbird*). This comparison will allow us to determine whether a compound (i.e., closed) structure encourages people to view the modified concept as being more distinct from the head noun concept than does a phrase-like structure, when the items are known compounds. On one hand, given that (in English), a closed compound structure is associated with more established compound words (see e.g., Kuperman and Bertram, 2013), if these effects are primarily driven by the notion that anything with a separate, existing, established compound name should have semantic differences from the head noun, one might expect that the open presentation would decrease the expectation of those semantic differences (and thus, an open presentation should attenuate the modification and inverse modification effects). On the other hand, simply presenting these well-known compound words with a space is, perhaps, unlikely to overcome the participants' knowledge that these are, in fact, established compound words. Thus, spacing might not be influential, because the participants might be unlikely to believe that the inserted space indicates a novel phrase.

Methods

Participants

Hundred and sixty three participants took part in the experiment. Each of the 12 lists was seen by a minimum of 13 and a maximum of 15 participants. We set a target of 15 participants per condition, but were not able to test the full number. We initially set a larger target per condition than in Experiment 1 because the design is more complicated. However, the effect sizes are quite large, and the sample size we obtained is more than sufficient.

Materials and Design

As in the previous experiment, we manipulated whether the concept was modified or unmodified, and the likelihood of the property (e.g., Some, Almost All, and Almost No). Modification and Likelihood were within-subject variables and were counterbalanced into 6 lists as in Experiment 1. The materials were identical to those used in Experiment 1. We also manipulated whether the modified concept was presented as a closed compound (e.g., *blackbird*) as in Experiment 1, or as an open compound (e.g., *black bird).* Spacing was a between-subjects factor to avoid drawing attention to this factor of interest and, thus, there were 12 lists (six with open items and six with closed items). Each participant saw one list. The design was a 2 (Modification) by 3 (Likelihood) by 2 (Spacing) crossed factorial design.

Procedure

The procedure was identical to Experiment 1.

Results and Discussion

Data Analysis

The descriptive statistics are shown in **Table 5**. We analyzed the data using LME regression models in which subject and item were entered as crossed random effects, and Modification (modified vs. unmodified), Likelihood (Almost All, Some, Almost No) and Spacing (open vs. closed) were entered as fixed effects, using the mixed and contrast commands in Stata (StataCorp, 2017). The *mixed* function outputs coefficients (i.e., estimates) for simple effects at the first level of the other categorical variables and at the mean of the other continuous variables in the model (see **Table 6**). For testing our hypotheses, these coefficients are not directly interpretable because they represent the simple effect of a variable at the first level of other variables. The relevant statistical tests for addressing our research questions concern interactions and simple effects, which are reported in the following text. The *contrast* function in Stata was used to conduct these analyses. We report the tests conducted on these fixed effects. Tests of simple effects (the effect of a factor at one level of another factor) were conducted to follow up on statistically significant interactions, because in the case of significant interactions, the main effects are not informative.

Results

As in Experiment 1, we found modification and inverse modification effects. There was a significant interaction between

TABLE 5 | Mean (SE) judged percentage of category members having the test property by level of Likelihood from Experiment 2.

		Percent of category members (*SE*)		
Spacing	**Condition**	**Almost All**	**Some**	**Almost No**
Closed	Unmodified	91.1 (0.24)	35.5 (0.41)	7.3 (0.25)
	Modified	64.9 (1.04)	27.3 (0.61)	11.0 (0.54)
Open	Unmodified	88.6 (0.37)	40.3 (0.47	8.8 (0.37)
	Modified	65.7 (0.97)	32.7 (0.68)	13.5 (0.57)

TABLE 6 | Experiment 2 Mixed Model Coefficients.

Variable	Coefficient	Standard Error	*z*	*P* > |z|
Spacing	0.8	1.2	0.7	0.481
Likelihood: Almost no	−53.9	0.81	−66.6	0.000
Likelihood: Some	−37.6	0.81	−46.5	0.000
Spacing × Likelihood Almost no	−1.7	1.13	1.5	0.130
Spacing × Likelihood Some	4.6	1.13	4.1	0.000
Modification	26.2	0.81	32.4	0.000
Spacing × Modification	−3.3	1.13	−2.9	0.003
Mod × Likelihood: Almost no	−29.9	1.14	−26.1	0.000
Mod × Likelihood: Some	−18.0	1.14	−15.8	0.000
Spacing × Mod × Likelihood: Almost no	2.2	1.59	1.4	0.162
Spacing × Mod × Likelihood: Some	2.7	1.59	1.7	0.097
Constant	64.9	0.94	69.1	0.000
Subjects	1.7	0.06		
Items	1.3	0.09		

Modification and Likelihood, $\chi^2(2) = 1314.37$, $p < 0.00001$. The tests of the simple effects revealed that the Almost All and Some conditions both led to a significant modification effect, $z = 43.48$, $p < 0.0001$ and $z = 13.87$, $p < 0.0001$, respectively, while the Almost No condition led to a significant inverse modification effect, $z = -7.53$, $p < 0.0001$. These effects were not affected by Spacing (e.g., *blackbird* vs. *black bird*); the three-way interaction between Spacing, Modification, and Likelihood was not significant, $\chi^2(2) = 3.19$, $p = 0.20$. Thus, in the case of known compounds, both the open and closed structure produced the same size of modification (or inverse modification) effects.

Although Spacing did not influence the two-way interaction between Modification and Likelihood (as indicated by the lack of three-way interaction), it did interact with Likelihood, $\chi^2(2) = 55.19$, $p < 0.0001$. In the Some condition, open items received higher ratings ($M = 36.5$, $SE = 0.83$) than did the closed items ($M = 31.4$, $SE = 0.85$), $z = 4.84$, $p < 0.0001$. However, spacing did not have an effect in the Almost All condition, $z < 1$. Influence of spacing in the Almost No condition was marginally significant, $z = 1.9$, $p = 0.06$, with ratings for open items being slightly higher ($M = 11.1$, $SE = 0.83$) than for closed items ($M = 9.11$, $SE = 0.85$). In sum, participants who received the open items gave higher ratings for the Some condition, whether the concept was modified or not, and, thus, this increase did not influence the modification effect itself.

Spacing also interacted with Modification, $\chi^2(1) = 6.62$, $p < 0.01$. As expected given that the compound was never seen in the unmodified condition, spacing did not affect the unmodified items, $z = 1.26$, $p = 0.21$. For the modified items, open items received higher ratings ($M = 45.7$, $SE = 0.80$) than did closed items ($M = 34.4$, $SE = 0.81$), $z = 2.93$, $p = 0.003$. To illustrate, participants were more willing to attribute an unknown property when the combined concept was expressed as an open form (*black bird*) than when it was a closed form (e.g., *blackbird*).

To further explore the influence of semantic transparency, as in Experiment 1, we included a *post hoc* examination of whether the similarity between the head and the compound influenced the ratings by including the LSA measure for the head and the compound in a model that also included Spacing and Likelihood (see **Table7**). As in Experiment 1, the analysis was restricted to only the modified concept condition because the compound was never seen in the unmodified condition and, thus, it would not be meaningful to include these items when examining the impact of the similarity between the compound and head. Spacing did not influence the nature of the interaction between LSA and Likelihood as indicated by the lack of interaction between these three variables, $\chi^2(2) = 1.24$, $p = 0.54$. As in Experiment 1, there was a significant interaction between the LSA measure and Likelihood, $\chi^2(2) = 73.64$, $p < 0.0001$. Further analysis indicated a significant slope for LSA in the Almost All and Some conditions, $z = 7.71$, $p < 0.0001$ and $z = 2.91$, $p < 0.004$, respectively, with the slope for transparency ratings being steeper in the Almost All condition than in the Some condition, $\chi^2(1) = 35.07$, $p < 0.0001$. However, the semantic association (as indicated by LSA) between the compound and the unmodified head did not affect the estimates in the Almost No condition, $z = 1.05$, $p = 0.29$. In sum, exactly as in Experiment 1, the modification effects were significantly smaller for compounds that were more semantically similar to their heads in the Almost all and Some conditions, but semantic similarity between the compound and the unmodified head had no effect in the Almost No condition. See **Figure 3**.

As in Experiment 1, to further examine whether property inference is influenced by degree of semantic transparency, we also used semantic transparency ratings by human participants from a database containing semantic transparency ratings for over 8000 compounds (Gagné et al., in press). The judgment of how much meaning of the head was retained in the compound was entered into a model that also included Spacing and Likelihood and was restricted to only the modified concept condition (see **Table 8**). Spacing did not influence the nature of the interaction between Transparency judgments and Likelihood as indicated by the lack of interaction between these three variables, $\chi^2(2) = 0.50$, $p = 0.78$. However, there was an interaction between transparency rating and Likelihood, $\chi^2(2) = 117.67$, $p < 0.0001$. Analysis of this interaction indicated a significant slope for the transparency rating in the Almost All and Some conditions, $z = 9.54$, $p < 0.0001$ and $z = 4.02$, $p < 0.001$, respectively, with the slope for transparency judgments being steeper in the Almost All condition than in the Some condition, $\chi^2(1) = 40.95$, $p < 0.0001$. As transparency increased, ratings for the properties increased for the modified items, meaning that more transparency would correspond to smaller modification effects. However, transparency judgments did not affect the estimates in the Almost No condition, $z = 0.31$, $p = 0.76$. See **Figure 4**.

TABLE 7 | Experiment 2 Mixed Model Coefficients with LSA: Modified condition only.

Variable	Coefficient	Standard Error	z	P > |z|
Spacing	1.7	2.2	0.8	0.418
Likelihood: Almost no	−47.0	1.7	−27.5	0.000
Likelihood: Some	−32.4	1.7	−18.9	0.000
Spacing × Likelihood Almost no	0.2	2.4	0.8	0.943
Spacing × Likelihood Some	4.0	2.4	1.7	0.094
LSA	42.8	6.0	7.1	0.000
Spacing × LSA	-5.5	5.9	−0.9	0.353
LSA × Likelihood: Almost no	−39.2	6.0	−6.6	0.000
LSA × Likelihood: Some	−28.0	6.0	−4.6	0.000
Spacing × LSA × Likelihood: Almost no	9.1	8.3	1.1	0.272
Spacing × LSA × Likelihood: Some	5.9	8.4	0.71	0.480
Constant	56.4	2.0	28.7	0.000
Subjects	2.1	0.07		
Items	1.9	0.10		

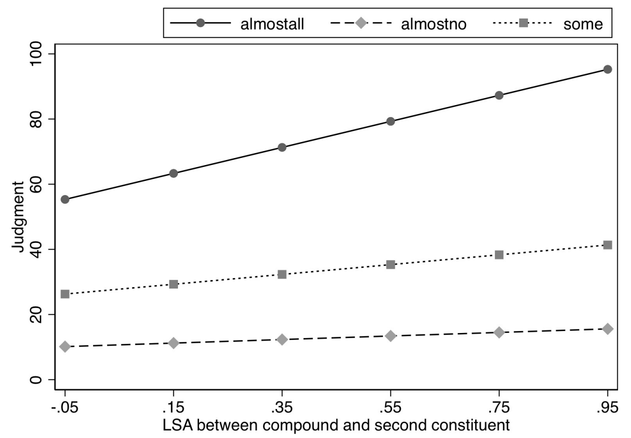


FIGURE 3 | Effect of semantic similarity (LSA) on judged likelihood of property in Experiment 2.

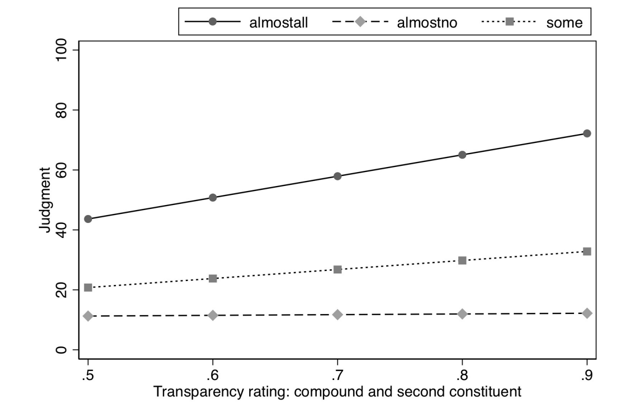


FIGURE 4 | Effect of semantic similarity (human ratings) on judged likelihood of property in Experiment 2.

Discussion

Experiment 2 replicated the effects observed in Experiment 1, including the effects of the semantic distance between the compound and the head. The effects of both LSA and semantic transparency ratings indicate that the modification effect gets larger as the semantic distance between compound and head increases, as one would expect. On the other hand, the inverse modification effect again seems immune to the effects of semantic distance between the compound and the head.

The manipulation of spacing had no impact on the modification or inverse modification effects with the materials in Experiment 2, nor did it affect the way in which the semantic distance interacted with the modification or inverse modification effect. Thus, although spacing can be an important visual cue of whether a compound word actually exists (versus being a novel phrase), when one actually knows of the existence of the compound, spacing is relatively unimportant in inferring properties of the compound.

TABLE 8 | Experiment 2 Mixed Model Coefficients with Semantic Transparency judgments: Modified condition only.

Variable	Coefficient	Standard Error	*z*	*P* > |z|
Spacing	4.8	7.3	0.7	0.511
Likelihood: Almost no	3.3	7.3	0.4	0.657
Likelihood: Some	−5.9	7.4	−0.8	0.424
Spacing × Likelihood Almost no	−2.2	10.2	−0.2	0.832
Spacing × Likelihood Some	7.0	10.3	0.7	0.493
Semantic Transparency	74.8	8.8	8.5	0.000
Spacing × ST	−6.7	9.1	−0.7	0.458
ST × Likelihood: Almost no	−72.6	9.2	−7.9	0.000
ST × Likelihood: Some	−40.4	9.3	−4.4	0.000
Spacing × ST × Likelihood: Almost no	6.9	12.8	0.54	0.590
Spacing × ST × Likelihood: Some	−1.6	12.9	−0.13	0.898
Constant	5.5	7.1	0.77	0.441
Subjects	2.2	0.07		
Items	1.8	0.10		

EXPERIMENT 3

Experiments 1 and 2 showed both modification and inverse modification effects for lexicalized compounds, with a strong asymmetry between the two effects. In addition, we found that the presentation of the compound word (open vs. closed), had no effect on the modification or inverse modification effects with these well-known compound words. This experiment investigates whether these effects depend on the specific knowledge of the known compound word, by replicating Experiment 2, except that we replaced the modifier of the compound words with a non-word (e.g., *blackbird* might become *flegbird*). In particular, the general principle that different names imply real underlying differences (Synonomy Avoidance, Carstairs-McCarthy, 2010; Principle of Contrast, Clark, 1993; and Mutual Exclusivity Principle, Markman, 1989) strongly suggests that manipulations that increase the likelihood of a name appearing to be permanent and unique should affect the extent to which participants infer properties from heads to compounds. The spacing manipulation failed to affect the modification and inverse modification effects in Experiment 2, however. It is possible that this failure is due to the fact that that these compound words are already well known. If so, then unknown words presented as closed compounds should lead to larger effects than unknown words presented as open compounds.

Methods

Participants

Seventy-two participants took part in the experiment. One participant was removed due to a computer problem during data collection. Experiments 1 and 2 had quite large effect sizes, so we set a lower target number of participants at 6 per condition.

Materials and Design

The head nouns from the previous experiments were used. Wuggy (Keuleers and Brysbaert, 2010) was used to generate non-words, which then replaced the first constituents of the compounds. The non-words were attached to the head nouns to create fake compounds with realistic compound structure (e.g., *blackbird* could become *flegbird*). The unknown predicates from the previous experiments were used and matched with the same head nouns. The materials were counterbalanced as in the previous experiments. As in Experiment 2, we manipulated whether the concept was modified or unmodified, the likelihood of the property (e.g., Some, Almost All, and Almost No), and whether the compound was presented as open or closed. Modification and Likelihood were within-subject variables and were counterbalanced into 6 lists as in Experiment 1. Spacing was a between-subjects factor to avoid drawing attention to this factor of interest and, thus, there were 12 lists (six with open items and six with closed items). Each participant saw one list. The design was a 2 (Modification) by 3 (Likelihood) by 2 (Spacing) crossed factorial design.

Procedure

The procedure was identical to the previous experiments. Because this experiment used non-word modifiers, additional instructions from previous work using non-word modifiers (Spalding and Gagné, 2015) were added. In addition to the task instructions from the previous studies, the participants were told: "When reading, people often come across unfamiliar words. One of our goals is to understand how people interpret phrases and compounds that contain such words. Therefore, in this experiment, some of the items will contain unfamiliar phrases (e.g., flug dogs)."

Results and Discussion

Data Analysis

The descriptive statistics are shown in **Table 9**. We analyzed the data using LME regression models in which subject and item were entered as crossed random effects, and Modification (modified vs. unmodified), Likelihood (Almost All, Some, Almost No) and Spacing (open vs. closed) were entered as fixed effects, using the mixed and contrast commands in Stata (StataCorp, 2017). The *mixed* function outputs coefficients (i.e., estimates) for simple effects at the first level of the other categorical variables and at the mean of the other continuous variables in the model (see **Table 10**). For testing our hypotheses, these coefficients are not directly interpretable because they represent the simple effect of a variable at the first level of other variables. The relevant statistical tests for addressing our research questions concern interactions and simple effects, which are reported in the following text. The *contrast* function in Stata was used to conduct these analyses. We report the tests conducted on these fixed effects. Tests of simple effects (the effect of a factor at one level of another factor) were conducted to follow up on statistically significant interactions, because in the case of significant interactions, the main effects are not informative.

Results

The three-way interaction between Spacing, Modification, and Likelihood was significant, $\chi^2(2) = 44.99$, $p < 0.0001$. To investigate the nature of this interaction, we then carried out separate analyses by level of Likelihood. There was a significant interaction between Modification and Spacing at each level of Likelihood, $\chi^2(1) = 65.3$, $p < 0.0001$; $\chi^2(1) = 4.15$, $p < 0.05$; and $\chi^2(1) = 4.17$, $p < 0.05$ at Almost All, Some, and Almost No, respectively. At each level of Likelihood, the interaction between Modification and Spacing indicated smaller modification (or inverse modification) effects when the item was presented with a space, compared to when it was presented as closed. Unlike Experiment 2, in this case, adding a space clearly attenuated the modification and inverse modification effects, as should be expected if the closed structure is seen by participants as more likely to indicate a permanent, unique name.

Nevertheless, as in the previous experiments, we consistently found modification and inverse modification effects. The tests of

TABLE 9 | Mean (SE) judged percentage of category members having the test property by level of Likelihood from Experiment 3.

		Percent of category members (*SE*)		
Spacing	**Condition**	**Almost All**	**Some**	**Almost No**
Closed	Unmodified	91.5 (1.3)	38.8 (2.6)	9.3 (2.5)
	Modified	74.7 (5.0)	33.6 (3.6)	14.1 (3.7)
Open	Unmodified	91.9 (1.7)	39.1 (2.8)	6.9 (1.7)
	Modified	85.7 (3.1)	36.2 (3.3)	9.4 (2.4)

TABLE 10 | Experiment 3 Mixed Model Coefficients.

Variable	Coefficient	Standard Error	z	P > |z|
Spacing	11.1	1.69	6.53	0.000
Likelihood: Almost no	−60.6	0.984	−61.1	0.000
Likelihood: Some	−41.1	0.984	−41.8	0.000
Spacing × Likelihood Almost no	−15.7	1.37	−11.5	0.000
Spacing × Likelihood Some	−8.3	1.37	−6.1	0.000
Modification	16.8	0.984	17.1	0.000
Spacing × Modification	−10.6	1.37	−7.7	0.000
Mod × Likelihood: Almost no	−21.6	1.39	−15.5	0.000
Mod × Likelihood: Some	−11.6	1.39	−8.4	0.000
Spacing × Mod × Likelihood: Almost no	12.9	1.94	6.6	0.000
Spacing × Mod × Likelihood: Some	8.2	1.94	4.2	0.000
Constant	74.6	1.22	61.5	0.000
Subjects	35.1	6.29		
Items	1.29e-09	5.15e-9		

the simple effects revealed that the Almost All conditions, both open and closed, led to significant modification effects, $z = 6.85$, $p < 0.0001$ and $z = 17.95$, $p < 0.0001$, respectively. The Some conditions, both open and closed, led to significant modification effects, $z = 3.59$, $p = 0.001$ and $z = 6.34$, $p < 0.0001$, respectively, while the Almost No conditions, both open and closed, led to significant inverse modification effects, $z = -3.16$, $p = 0.003$ and $z = -5.92$, $p < 0.0001$, respectively.

Discussion

We once again replicate the robust modification and inverse modification effects, even when the modifiers of the compounds used in the previous experiments are replaced with non-words. Thus, although the previous experiments showed that the semantics of the known compounds contribute to the modification effects, it is clear that the main aspects of the effects are maintained even in cases where no known semantics can be brought to bear on the inference. In addition, unlike Experiment 2, we found that adding a space attenuates both the modification and the inverse modification effect. Presumably, when the participants do not have semantic knowledge to fall back on, they make more use of the visual cue given by the spacing to indicate the permanence of the compound.

GENERAL DISCUSSION

Across all experiments, we found very robust modification (in the Almost All and Some condition) and inverse modification effects (in the Almost No condition) and these effects were observed for both closed and open structures. Thus, although people appear to infer previously unknown properties from heads to compounds, in accordance with a general principle of categorical inference, they make those inferences in line with the general principle that unique names imply underlying semantic differences (as suggested by Synonomy Avoidance, Carstairs-McCarthy, 2010; Principle of Contrast, Clark, 1993; and Mutual Exclusivity Principle, Markman, 1989). In short, when people infer properties from heads to compounds, they do so by coordinating these two general principles regulating the relationship between categories and sub-categories.

In addition, we found that there are two important asymmetries between the modification and inverse modification effects: First, the modification effects are numerically much larger than the inverse modification effects. Second, the size of the modification effect is quite sensitive to the existing semantic distance between a compound word and its head, while the inverse modification effect appears to be entirely insensitive to the existing semantic distance. These results suggest that in making categorical inferences, people are sensitive to the fact that there is an important difference between properties generally true of a category and properties generally untrue of a category, as suggested by previous work on concepts (e.g., Murphy and Medin, 1985). Because our concepts are generally organized around properties that are believed to be true, rather than false, of those concepts, the modification effect (involving properties generally true of the concepts) appears to be much more tightly tied to the existing semantics of the known compounds. In particular, the modification effect seems to be enhanced by the existing differences (when dealing with known compound words). In short, for the modification effect, the more "semantically modified" the compound is, relative to the head, the larger the effect. This result is also consistent with the observation that the more modifiers that are included, the larger the modification effect becomes (Connolly et al., 2007); for example, the modification effect was larger for *Baby Peruvian ducks have webbed feet* than for *Baby ducks have webbed feet*.

The inverse modification effect, on the other hand, seems to be a kind of base line effect that reflects just the general principle that a unique name implies some underlying difference, and seems to be entirely insensitive to the known semantic difference between a compound and its head. Yet, it is sensitive to spacing (in Experiment 3, where the materials are not known compounds), suggesting that the inverse modification effect is sensitive to the likelihood that the "compound" is a permanent, existing word, just not to the semantics of that word.

We have suggested that this pattern of results is consistent with the idea that the results are driven by an underlying conceptual difference between properties considered true of a concept and those considered false. In short, the semantic change that accompanies modification of a concept should be more likely to affect properties true of the head than properties false of the head, on average. We are not, of course, assuming that there are no properties that are true of the compound but false of the head (a well-known example is that pet fish often live in glass bowls, but fish do not usually live in glass bowls). Rather, our point is that there are very many things that are not thought of as true of the head concept (and, indeed, are unrelated to the head) and most of them will remain unrelated to the compound (so, neither fish nor pet fish enjoy salsa dancing, dissolve plastic, or trap dust mites). Thus, any individual, unknown feature that is presented as false of the head, has a relatively strong likelihood of being false of the compound.

Furthermore, features thought of as true of the head concept are often related to each other (see, e.g., Murphy and Medin, 1985), such that a modification of that head that affects one of the features is likely to affect others. For example, if we think of, say, wings, feathers, and flying as being commonly true of birds, we find that a modification that affects one, often affects the others (e.g., birds that cannot fly generally have unusually small wings relative to body size, and their feathers are often quite different from what we think of as "normal" bird feathers—penguin feathers or ostrich feathers, for example). On the other hand, take three things (somewhat randomly) not thought of as true of birds, say explosive, earned a PhD, and made of glass. Now, any of these things might be true of a particular bird in some particular circumstance (a bird used to deliver a bomb strapped to it, a bird given an honorary degree after years of use in a laboratory, or a glass statue of a bird). However, having a compound that affects one of those features (say, an exploding bird), is unlikely to have much consequence for the others; the exploding bird is not particularly likely either to have earned a PhD or to be made of glass.

The point is that properties that are clearly not true of a concept are likely to be very far from the region of semantic space which the head and the compound inhabit, and hence to be unaffected by the relatively small movement in semantic space normally associated with the difference between the head and the compound. In addition, properties that share only the fact that they are false of some concept are likely to be drawn from much more distinct semantic spaces than properties that are all true of that same concept, such that changes to one false property are unlikely to have consequences for other false properties, compared to true properties. Hence, the semantic contrast created by a known compound might be less likely to strongly influence things thought of as false of the head, compared to things thought of as true of the head.

Finally, we found that the presentation format (open or closed) had no effect on the size of the modification and inverse modification effects when the materials were known compound words, but when the modifier of those known compounds was replaced with a non-word, the inclusion of a space attenuated both the modification and inverse modification effects. In general, it seems that when a compound is known to exist, the presentation format does not affect property inference. This is quite reasonable, if the presentation format functions primarily to indicate the higher likelihood of existence as a separate, unique name (when there is no space), because when the compounds are known to exist, the lack of a space does not add to the participants' certainty that this is a unique name. However, when the modifier of the compound is replaced by a non-word, having the closed structure makes it more likely (in the participants' view) that the letter string is intended to reflect a permanent, unique name, and hence the modification and inverse modification effects are larger.

An alternative explanation that might occur to the reader is that modification just increases uncertainty, and this explains the modification effects, such that, in essence, participants are simply less likely to use the extremes of the scale. In this view, it is the uncertainty that causes the modification effects (see, e.g., suggestions by Jönsson and Hampton, 2008, 2012; Hampton et al., 2011) rather than the coordination of general principles of categorical inference and of unique names implying underlying semantic differences (Synonomy Avoidance, Carstairs-McCarthy, 2010; Principle of Contrast, Clark, 1993; and Mutual Exclusivity Principle, Markman, 1989). However, uncertainty, though superficially appealing as an explanation, does not explain the details of the effects. First, several aspects of uncertainty were investigated with respect to modification effects in novel compounds, and were found not to account for the effects (Gagné and Spalding, 2014b). Second, presumably, participants would be far more uncertain about properties with respect to novel compounds than known compounds, so an uncertainty explanation predicts that the effects would be larger for novel compounds than for known compounds, but in fact the effects with known compounds are much larger than with novel compounds (e.g., the size of the effects in the current experiments compared to those in Spalding and Gagné, 2015). With respect just to the experiments in the current paper, uncertainty would be far higher with non-word modifiers, so the effects should be much larger with non-word modifiers if those effects are driven by uncertainty, but again this is not the case. Third, uncertainty would be higher for false features (as the relations of such properties even to the head are very unclear, i.e., uncertain), so the effects for the false properties should be larger than for the true properties, but again this is the reverse of the case, both with known compounds in the current experiments and with novel compounds in Spalding and Gagné (2015). Finally, avoiding extreme values on the scale would not explain the modification effect in the "some" condition of the current experiments. Thus, uncertainty seems not to be a good explanation for the overall pattern of the results.

The current results add to the literature on modification effects by showing that the process of property inference is highly consistent across both novel modifications (i.e., conceptual combination) and known compound words, though there are some differences relating to the extent to which the modification is considered to be permanent and unique, and if it is permanent and unique, the extent to which the known compound differs, semantically, from the head. Thus, people know that the purpose of modification is to signal some underlying differences from the unmodified item. The more stable and permanent that modification is believed to be, the more strongly those underlying differences are signaled. Importantly, the modification and inverse modification effects, whether with novel or known compounds, appear to result from inference processes, rather than from a direct, conceptually-driven property inheritance process in which properties of a head are automatically linked to any new modification of that head (see, e.g., Gagné and Spalding, 2011, 2014b, 2015; Spalding and Gagné, 2015, for discussion of this theoretical distinction).

The current results also add to our understanding of the general principle that a unique name implies real underlying differences. An interesting question about the principles of Synonomy Avoidance (Carstairs-McCarthy, 2010), Contrast (Clark, 1993), and Mutual Exclusivity (Markman, 1989) is, what happens once an underlying difference has been identified? Do the principles require simply a minimal difference for each unique name? That is, does knowing of the existence of an underlying difference fully justify the existence of a unique name? If so, then one would expect that items that are already known to differ substantially should be less likely to result in modification (or inverse modification) effects. If there is already a known difference, one need not infer other differences in order to fulfill the needs of Synonomy Avoidance, or Contrast, or Mutual Exclusivity—the compound is already known to differ from the head, so there is no need to create further differences via property inference. Our results strongly suggest that this is not the case (see also Spalding and Gagné, 2015, Experiment 3). Instead, the more the compound is already known to differ from the head, the more it is expected to differ with respect to new, unknown properties, such as those being tested in our experiments. Thus, although the general principle that unique names imply real underlying differences appears to strongly affect property inferences, it is not the case that minimal Synonomy Avoidance, or Contrast, or Mutual Exclusivity is what people are expecting as the result of a unique name. On the other hand, it is also not the case that the unique name implies a need to maximally contrast, such that properties are not inferred at all from the unmodified to the modified. Instead, people appear, generally, to infer properties from the unmodified to the modified noun in an inverse relationship with the degree of contrast that they expect, based on what they know of the modified item, on other generally useful cues such as spacing (when appropriate) or number of modifiers, and on the nature of the property to be inferred (generally true or generally false of the unmodified item).

Conclusion

This series of experiments indicates that people infer new properties from unmodified nouns to compounds with that noun as the head in accordance with the principle of categorical inference, but also in accordance with the general principle that a unique name implies underlying semantic differences. They appear to make these inferences not in an automatic or mechanical way, but by using the information that they have available to them about the meaning of the compound, the nature of the property (true or false of the head), as well as other cues that they believe are likely to be related to the extent to which the unique name is well established and permanent. Finally, these experiments indicate that property inference follows the same principles, regardless of whether the compound is novel or well known, though the extent to which the compound is believed to be established does affect the degree to which the property is likely to be inferred.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the Research Ethics Office of the University of Alberta with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Research Ethics Office of the University of Alberta.

AUTHOR CONTRIBUTIONS

TS and CG contributed to the theoretical framing, writing, data analysis, and revision. GL contributed to theoretical framing and revision. KN and JC contributed to the data collection and revision. All authors provided critical feedback and helped to shape the research and the manuscript.

FUNDING

This research was supported by the Natural Science Engineering Research Council of Canada Discovery Grants (250028 and 05100) to TS and CG.

REFERENCES

Baldwin, T., Martinez, D., Penman, R. B., Kim, S. N., Lui, M., Wang, L., et al. (2010). "Intelligent linux information access by data mining: the ILIAD project," in *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, Los Angeles, CA.

Carstairs-McCarthy, A. (2010). *The Evolution of Morphology*. Oxford: Oxford University Press.

Clark, E. V. (1993). *The Lexicon in Acquisition*. Cambridge: Cambridge University Press.

Connolly, A. C., Fodor, J. A., Gleitman, L. R., and Gleitman, H. (2007). Why stereotypes don't even make good defaults. *Cognition* 103, 1–22.

Dragos, V., Battistelli, D., and Kelodjoue, E. (2018). "Beyond sentiment and opinions: exploring social media with appraisal categories," in *Proceedings of the 21st International Conference on Information Fusion (FUSION)*, Cambridge.

Gagné, C. L., and Murphy, G. L. (1996). Influence of discourse context on feature availability in conceptual combination. *Discourse Process.* 22, 79–101.

Gagné, C. L., and Spalding, T. L. (2009). Constituent integration during the processing of compound words: does it involve the use of relational structures? *J. Mem. Lang.* 60, 20–35.

Gagné, C. L., and Spalding, T. L. (2011). Inferential processing and meta-knowledge as the bases for property attribution in combined concepts. *J. Mem. Lang.* 65, 176–192.

Gagné, C. L. and Spalding, T. L. (2014a). "Conceptual composition: the role of relational competition in the comprehension of modifier-noun phrases and noun-noun compounds," in *The Psychology of Learning and Motivation*, ed B. H. Ross (Amsterdam: Elsevier), 97–130

Gagné C. L., and Spalding T. L. (2014b). Subcategorisation, not uncertainty, drives the modification effect. *Lang. Cogn. Neurosci.* 29, 1283–1294. doi: 10.1080/23273798.2014.911924

Gagné, C. L., and Spalding, T. L. (2015). "Semantics, concepts, and meta-cognition: attributing properties and meanings to complex concepts," in *Semantics of Complex Words* eds L. Bauer and L. Körtvélyessy and P. Stekauer (Berlin: Springer), 9–25

Gagné, C., Spalding, T. L. and Schmidtke, D. (in press). LADEC: large database of english compounds. *Behav. Res. Methods* doi: 10.7939/r3-dyqx-9b36

Gagné, C. L., Spalding, T. L., and Kostelecky, M. (2017). "Conceptual combination, property inclusion, and the artistotelian-thomistic view of concepts," in *Compositionality and Concepts in Linguistics and Psychology*, eds J. Hampton and Y. Winter (Berlin: Springer), 223–244.

Hampton, J. A. (1987). Inheritance of attributes in natural concept conjunctions. *Mem. Cogn.* 15, 55–71.

Hampton, J. A. (1991). "The combination of prototype concepts," in *The Psychology of Word Meanings*, ed P. J. Schwanenflugal (Hillsdale, NJ: Erlbaum).

Hampton, J. A., Passanisi, A., and Jönsson, M. L. (2011). The modifier effect and property mutability. *J. Mem. Lang.* 64, 233–248. doi: 10.1016/j.jml.2010.12.001

Hirschman, L., and Gaizauskas, R. (2001). Natural language question answering: the view from here. *Nat. Lang. Eng.* 7, 275–300.

Ji, H., Gagne, C. L., and Spalding, T. L. (2011). Benefits and costs of lexical decomposition and semantic integration during the processing of transparent and opaque English compounds. *J. Mem. Lang.* 65, 406–430.

Jönsson, M. L., and Hampton, J. A. (2008). On prototypes as defaults (comment on Connolly, Fodor, Gleitman, and Gleitman, 2007). *Cognition* 106, 913–923.

Jönsson, M. L., and Hampton, J. A. (2012). The modifier effect in within-category induction: default inheritance in complex noun phrases. *Lang. Cogn. Process.* 27, 90–116. doi: 10.1080/01690965.2010.544107

Keuleers, E., and Brysbaert, M. (2010). Wuggy: a multilingual pseudoword generator. *Behav. Res. Methods* 42, 627–633. doi: 10.3758/BRM.42.3.627

Kuperman, V., and Bertram, R. (2013). Moving spaces: spelling alternation in English noun-noun compounds. *Lang. Cogn. Process.* 28, 939–966.

Landauer, T. K., and Dumais, S. T. (1997). A solution to plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* 104:211.

Libben, G. (1998). Semantic transparency in the processing of compounds: consequences for representation, processing, and impairment. *Brain Lang.* 61, 30–44.

Libben, G., and Jarema, G. eds (2006). *The Representation and Processing of Compound Words*. Oxford: Oxford University Press.

Lieber, R. (2004). *Morphology and Lexical Semantics*. Cambridge: Cambridge University Press.

Lieber, R. (2009). "A lexical semantic approach to compounding," in *The Oxford Handbook of Compounding*, ed. R. Lieber and P. Štekauer (Oxford: Oxford University Press).

Markman, E. M. (1989). *Categorization and Naming in Children: Problems of Induction*. Cambridge, MA: MIT Press.

Maynard D., and Funk A. (2012). "Automatic detection of political opinions in tweets," in: *The Semantic Web: ESWC 2011 Workshops. ESWC 2011. Lecture Notes in Computer Science*, eds García-Castro R., Fensel D., Antoniou G (Berlin: Springer)

Murphy, G. L. (2002). *The Big Book of Concepts*. Cambridge, MA: MIT Press.

Murphy, G. L., and Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychol. Rev.* 92, 289–316.

Osherson, D. N., Smith, E. E., Wilkie, O., López, A., and Shafir, E. (1990). Category-based induction. *Psychol. Rev.* 97, 185–200.

Spalding, T. L., and Gagné, C. L. (2015). Property attribution in combined concepts. *J. Exp. Psychol. Learn. Mem. Cogn.* 41, 693–707. doi: 10.1037/xlm0000085

Springer, K., and Murphy, G. L. (1992). Feature availability in conceptual combination. *Psychol. Sci.* 3, 111–117.

StataCorp (2017). *Stata Statistical Software: Release 15. College Station*. Texas, TX: StataCorp

Wang, L., Kim, S. N., and Baldwin, T. (2010). "Thread-level analysis over technical user forum data," in *Proceedings of the Australasian Language Technology Association Workshop*, Melbourne.

Wang, L., Lui, M., Kim, S. N., Nivre, J., and Baldwin, T. (2011). "Predicting thread discourse structure over technical web forums," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*, Edinburgh.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

TABLE A1 | Experimental materials.

Compounds	Predicates
Afterlives	Are dinural
Armpits	Have star shaped carpels
Arrowheads	Have large epiphyseal plates
Barbershops	Use POS software
Bathtubs	Are alloyed with cholorargyrite minerals
Billfolds	Contain casein proteins
Blackbirds	Require graminoids in their diet
Bloodstreams	Have geometric gradients
Blueberries	Contain high amounts of citrulline
Boyfriends	Exhibit dialectic social bonds
Brainwaves	Transmit delta oscillating networks
Bridegrooms	Are studied by malacologists
Broadcasts	Have screens made of polyamid fibers
Busybodies	Have a mobile quadrate bone
Candlesticks	Are used in making Nocello wine
Cheekbones	Secrete the chemical mercaptan
Clamshells	Contain mycoban
Classmates	Are agentic
Cloverleaves	Contain cis-3-hexenal
Coalmines	Are produced with neotame
Corkscrews	Are made from nickel antigorite
Courtyards	Consist of low, tight swards
Crosswords	Are made with combinations of graphemes
Deathtraps	Have a Weberian apparatus
Drawbridges	Have vectualic valves
Dustpans	Contain traces of mixita silica
Fairytales	Come from the Aleutian Islands
Fenceposts	Are built with pneumatic pounders
Fingertips	Are colored by diazo dyes
Floorshows	Became more common when the MFA was bolished
Flowcharts	Represent aeronautical data
Flowerpots	Are cooled in annealing ovens
Footprints	Are lithograph images
Gardenpaths	Have chester webbing
Gemstones	Contain alliin lyase
Goldfinches	Eat Danio rerio
Greyhounds	Have plantigrade paws
Gunpowder	Solidifies after the process of metasomatism
Hailstorms	Contain aromatic hydrocarbons
Hairnets	Are made from jute fibers
Heartburns	Are caused by high concentrations of oleic acid
Hovercrafts	Are facultatively ammonotelic
Inkstands	Have wheels made of cis-1,4-polyisoprene
Jellybeans	Are flavored with strobili
Keyholes	Are caused by structuram dynamics
Lampshades	Are formed from dies with infuse images

(Continued)

TABLE A1 | Continued

Compounds	Predicates
Loveboats	Run by compressing the chemical R134a
Madhouses	Follow plazzo hierarchies
Matchboxes	Are lacquered using phenolic compounds
Mincemeat	Has volumetric muscles
Molehills	Have pterylae growths
Moonbeams	Are tetrachromatic
Mothballs	Were first used in episkyros
Necklaces	Have pecten oculi
Notepaper	Is created using cellulose pulp
Paintbrushes	Are susceptible to thrips
Patchwork	Is artisanal
Pawnbrokers	Are descended from mouflons
Photocopiers	Have magnetic solenoids
Placemats	Are made from vulcanized rubber
Plywoods	Contain vitamin HB5
Poppyseeds	Produce purified pectinase
Quicksand	Is made of SiO2
Racehorses	Use vibrissae hairs for tactile sensation
Ribcages	Are secured with helical pieces
Rockpools	Are found in seiches
Rosebuds	Contain a cyanogenic seed
Scarecrows	Are of the genus Loriculus
Scrapbooks	Use the rules of boustrophedon
Seafood	Has cemetrum cells
Sketchpads	Contain protein polymers
Snakeskin	Is made using gliadin proteins
Soupspoons	Were first used in Sardinia
Spacesuits	Come from Witwatersand
Spyglasses	Are used as lorgnettes
Starfish	Swim in a state of tonic immobility
Stomachaches	Are caused by spermaceti
Sugarcanes	Are used by rabologists
Surfboards	Have wooden nocks
Thumbnails	Are sharpened using unite stones
Thunderbolts	Are applied with hexagonal torque
Timetables	Are made of creosote preserved wood
Tinfoil	Is developed using the calotype process
Toothpicks	Have CYL2 materials
Topsoil	Is a part of the pedosphere
Wastebaskets	Are made from baleen
Watermelons	Contain organic apiol
Wheelchairs	Are of ecclesiastical origin
Windpipes	Follow the SMLS structure
Wirecutters	Burn Bitumen based fuels
Witchdoctors	Are artiodactyles
Wristwatches	Have octave key saddles
Airplanes	Use KC-135 quantas
Campfires	Release the chemical HCL
Cavemen	Are of the heterogametic sex
Earthworms	Function with a hydrostatic skeleton

frontiers in Communication

ORIGINAL RESEARCH
published: 31 July 2019
doi: 10.3389/fcomm.2019.00031

Check for updates

Effects of Orthographic Forms on the Acquisition of Novel Spoken Words in a Second Language

Tania Cerni[1]*, *Bene Bassetti*[2] *and Jackie Masterson*[3]

[1] *Department of Psychology and Cognitive Science, University of Trento, Rovereto, Italy,* [2] *School of Education, University of Birmingham, Birmingham, United Kingdom,* [3] *Department of Psychology and Human Development, UCL Institute of Education, London, United Kingdom*

The orthographic forms of words (spellings) can affect word production in speakers of second languages. This study tested whether presenting orthographic forms during L2 word learning can lead speakers to learn non-nativelike phonological forms of L2 words, as reflected in production and metalinguistic awareness. Italian_{L1} learners of English as a Second Language (English_{L2}) were exposed to English_{L2} novel spoken words (pseudowords) and real words in association with pictures either from auditory input only (Phonology group), or from both auditory and orthographic input (Phonology & Orthography group, both groups $n = 24$). Pseudowords and words were designed to obtain 30 semi-minimal pairs, each consisting of a word and a pseudoword that contained the same target consonant, spelled with one letter or with double letters. In Italian double consonant letters represent a long consonant, whereas the English language does not contrast short and long consonants. After the learning phase, participants performed a production task (picture naming), a metalinguistic awareness task (rhyme judgment) and a spelling task. Results showed that the Phonology & Orthography group produced the same consonant as longer in double-letter than in single-letter lexical items, while this was not the case for the Phonology group. The former group also rejected spoken rhymes that contained the same consonant spelled with a single letter in one word and double letters in the other, because they considered these as two different phonological categories. Finally, the Phonology & Orthography group learned more novel words than the Phonology group, showing that orthographic input results in more word learning, in line with previous findings from native speakers.

Keywords: word learning, orthographic effects, second language, language production, metalinguistic awareness

OPEN ACCESS

Edited by:
Eva Kehayia,
McGill University, Canada

Reviewed by:
Marilyn Vihman,
University of York, United Kingdom
Caicai Zhang,
Hong Kong Polytechnic University, Hong Kong

Correspondence:
Tania Cerni
tania.cerni@gmail.com

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *31 October 2018*
Accepted: *21 June 2019*
Published: *31 July 2019*

Citation:
Cerni T, Bassetti B and Masterson J (2019) Effects of Orthographic Forms on the Acquisition of Novel Spoken Words in a Second Language. Front. Commun. 4:31. doi: 10.3389/fcomm.2019.00031

INTRODUCTION

Orthographic input impacts spoken word learning in native, second and novel languages (languages unknown to the participant, including artificial languages), although effects are different. In a native language, those who are exposed to both auditory and orthographic input learn more spoken words than those exposed only to auditory input. This positive effect of orthographic input – sometimes referred to as "orthographic facilitation" (Chambré et al., 2017)—was found in both children (Ehri and Wilce, 1979; Rosenthal and Ehri, 2008; Ricketts et al., 2009) and adults (Nelson et al., 2005; Miles et al., 2016), who memorized the spoken form and meaning of words more easily if they saw the word's written form. The same was found in child learners of English as a foreign language (Hu, 2008; Vadasy and Sanders, 2015).

Among adult second language learners, orthographic input facilitates the perception of the phonological form of novel L2 words. Thus, Escudero et al. (2008) found that Dutch_{L1} speakers of $\text{English}_{\text{L2}}$ could better recognize spoken $\text{English}_{\text{L2}}$ pseudowords containing one of the two vowels /æ/–/ε/, a confusable contrast for Dutch_{L1} speakers, when in a previous word learning phase the phonological forms of these words were accompanied by their orthographic forms (spellings). However, Escudero et al. (2014) found that the effects of orthographic input depend on the correspondences between the graphemes and phonemes in the native language, such that Dutch_{L2} learners were facilitated in Dutch word learning if the word's orthographic form contained grapheme-phoneme correspondences congruent with those in the native language, and inhibited when the two languages' grapheme-phoneme correspondences were incongruent.

In novel languages, the effects of orthographic input on the learning of the phonological forms of words are less clear-cut, ranging from positive to negative to no effects, depending on the difficulty of the sounds to be learned and familiarity of the symbols used to represent them. The first study to investigate the phenomenon (Erdener and Burnham, 2005) found positive effects in $\text{English}_{\text{L1}}$ and $\text{Turkish}_{\text{L1}}$ learners of two novel languages (Irish and Spanish). In particular, orthographic input facilitated the learning of phonological forms in native speakers of the phonologically transparent Turkish language, and learners of the phonologically transparent Spanish as a novel language. While all the languages studied in Erdener and Burnham's (2005) paper were written with the roman alphabet, Showalter et al. (2013) found that even unfamiliar orthographic symbols could improve the perception of spoken words in a novel language. In their study, English native speakers with no experience of Chinese were taught to associate pictures with Chinese pseudowords which varied in lexical tone. Participants who had learned the new spoken words together with their orthographic form in romanised Chinese outperformed those who had only received auditory input in a subsequent word recognition task. However, orthographic input does not seem to help novel language learners when the contrast is difficult. Showalter and Hayes-Harb (2015) carried out a study where $\text{English}_{\text{L1}}$ speakers with no knowledge of the Arabic language learned a set of Arabic pseudoword minimal pairs created to test their perception of the Arabic velar–uvular contrast (e.g., [kubu], [qubu]), which is particularly difficult for $\text{English}_{\text{L1}}$ speakers. Each word was associated with an image, and with either the word's orthographic form (experimental condition) or an unrelated string of Arabic symbols (control condition). Orthographic input did not facilitate the learning of this contrast, even when participants were instructed about the Arabic writing system, or when Arabic symbols were replaced with romanisation. Pytlyk (2011) also found that presenting orthographic input during word learning did not help Chinese phoneme discrimination in $\text{English}_{\text{L1}}$ speakers with no previous exposure to Chinese. There were no differences between presenting Chinese transcriptions written in the familiar roman alphabet or in an unfamiliar Chinese transcription system. Mathieu (2016) even found a negative effect of orthographic input presentation in $\text{English}_{\text{L1}}$ speakers acquiring a consonantal contrast in Arabic either with the spoken forms alone or else spoken forms plus orthographic forms in one of three scripts—Arabic, cyrillic and a hybrid roman/cyrillic script. Participants who had learned the spoken words with the orthographic forms were less accurate in recovering the target contrast than participants who had only heard auditory input, because familiar scripts had effects when the same grapheme represented a different sound in the native and the novel language. It appears that orthographic input affects the learning of the phonological forms of new words and sounds, with generally positive effects in L2 speakers and more varied effects in learners of novel languages.

Looking at the production of familiar L2 words by experienced L2 speakers, the words' orthographic forms can lead to non-nativelike production (Bassetti, 2007; Hayes-Harb et al., 2010). Effects can include sound addition, as when L2 speakers add an epenthetic sound corresponding to a so-called "silent letter," such as a [l] in *walk* (Bassetti and Atkinson, 2015). However, the most studied orthographic effects on L2 speech production happen because L2 speakers recode L2 orthographic forms using L1 grapheme-phoneme correspondences, which results in sound substitution, for instance when L2 speakers of American English pronounce a [t] or a [d] in words spelled with letters <t> or <d>, which native speakers produce as flaps (Vokic, 2011). Research by Bassetti and colleagues in particular shows that L2 orthographic forms can lead L2 speakers to establish a phonological contrast in their L2 phonological system that does not exist in the phonological system of the target language. Bassetti (2017) found that Italian $\text{English}_{\text{L2}}$ learners, whose native orthography uses double consonant letters to represent geminates (long consonants), tend to produce the same $\text{English}_{\text{L2}}$ consonant as longer when spelled with double consonants than with singleton consonants, for instance producing a longer /t/ in *kitty* than in *city*. This was found in a reading aloud task but also in a delayed word repetition task with or without orthographic input presented before the repetition. The fact that the effect appeared also when the spelling of the word was not available could attest to an orthography-influenced L2 phonological representation, or to the activation of the L2 orthographic representation during L2 speech production. In support of this interpretation, Italian speakers of $\text{English}_{\text{L2}}$ were found to produce minimal pairs distinguished by a geminate or singleton consonant such as /ˈfinːiʃ/—/ˈfiniʃ/ (*Finnish-finish*, both /ˈfɪnɪʃ/ in British English; Bassetti et al., 2018). Furthermore, they rejected rhymes containing the same consonant written with a singleton letter or with double letters (e.g., *very-merry*) because they considered these two phonological categories, a singleton and a geminate consonant, as in $\text{Italian}_{\text{L1}}$ (Bassetti et al., under revision).

Previous research has found, then, that orthographic forms can affect the perception and production of L2 words. However, these effects could be due to repeated exposures rather than being established at the point of first learning words. In the present study we aimed to investigate whether orthographic effects are found in the very early stages of acquiring vocabulary items. We investigated whether Italian_{L1} speakers of English_{L2} would learn a spoken novel word in English as containing a geminate (long) consonant if the word was presented with an orthographic form that contains double consonant letters. This outcome could come about as the result of recoding the English word's orthographic form according to Italian_{L1}'s grapheme-phoneme correspondence rules. The recoded orthographic form would then interact with the phonological form from the L2 spoken input—which contains a short consonant—resulting in a conflict between the two forms—and possibly a perceptual illusion—and result in a phonological representation containing a long consonant.

To test this, we compared the learning of pseudowords in two groups of Italian English_{L2} learners. One group received only phonological input during the learning phase (Phonology group) and the other received orthographic input simultaneously with the phonological input (Phonology & Orthography group). We then assessed participants' production, metalinguistic awareness and spelling of the target sounds. Items for the learning task comprised word and pseudoword pairs. The pseudoword in each pair was created by replacing the onset (initial consonant or consonant cluster) of the word (e.g., /ˈprɪnɪʃ/–/ˈfɪnɪʃ/, *prinish-finish*).

The first aim of the study was to assess effects of orthography on the spoken production of novel words. To address this aim, participants learned spoken stimuli in association with pictures. We then used a picture naming task to compare the target consonant duration ratio in the Phonology and Phonology & Orthography groups. Target consonant duration ratio was obtained by dividing the duration of the target consonant in the first word of the pair by the duration of same target consonant in the other word, a procedure that has been used in previous studies (Bassetti et al., 2018; Bassetti et al., under revision). There were three types of pairs: CC_{pw}-C_w pairs consisted of a pseudoword where the target consonant was spelled with double letters and a word where it was spelled with a single letter, such as *prinnish-finish*; CC_w-C_{pw} pairs consisted of a word where the target consonant was spelled with double letters and a pseudoword where it was spelled with a single consonant letter, such as *Finnish-prinish*; C_{pw}-C_w pairs consisted of a pseudoword and a word both containing the target consonant spelled with a single letter (e.g., *prinish-finish*). We had different predictions for each group for each type of pair, as follows. With regards to the Phonology & Orthography group, we predicted that they would have high ratios in pairs where pseudowords were spelled with double consonants and words with a single consonant (CC_{pw}-C_w pairs, e.g., *prinnish-finish*), because they were expected to produce responses where the consonant in the pseudoword was longer than the consonant in matched words. For this group, we predicted high ratios also in pairs consisting of a word spelled with double consonants and a pseudoword spelled with a single consonant (CC_w-C_{pw} pairs, e.g., *Finnish-prinish*), because we predicted that they would produce responses where the consonant in the word was longer than the consonant in the matched pseudoword. Finally, we predicted a ratio of around one for pairs where both pseudoword and word were spelled with a single consonant letter (C_{pw}-C_w pairs, e.g., *prinish-finish*), as both consonants would be produced with similar lengths.

A related aim was to assess how Italians would categorize and learn consonants in pseudowords in the absence of orthography. We had two different sets of predictions for the speech production of the Phonology group. If, in the absence of orthographic input, Italians categorize all English consonants as singletons, we predicted a ratio of around one for CC_{pw}-C_w and C_{pw}-C_w pairs, as both pseudowords and words spelled with a single letter would be produced as short, and we predicted a high ratio for CC_w-C_{pw} pairs, as the word spelled with double consonants would be produced as longer than the pseudoword. On the other hand, Bassetti (2017) argued that Italians may perceive the duration of English consonants as being on the boundary between short and long consonants, and perceive and produce such consonants as long or short depending on their spelling. If this hypothesis is correct, then in the absence of orthographic information Italians should categorize English consonants as either singleton or geminate. In line with this hypothesis, the Phonology group should produce all pairs with ratios higher than one because all pseudowords would be variably produced as either long or short consonants. The Phonology group would therefore have a lower ratio than the Phonology & Orthography group in CC_{pw}-C_w pairs, as the Orthography group would produce these consistently with a long consonant in the pseudoword and a short consonant in the word, whereas the Phonology group would produce some pseudowords with a long consonant and some with a short consonant. With C_{pw}-C_w pairs, the Phonology group would have a higher ratio than the Phonology & Orthography group's ratio of one. Finally, looking at CC_w-C_{pw} pairs, the Phonology group would have a lower ratio than the Phonology & Orthography group, because the latter would produce all pseudowords with short consonants, but the former would produce some with a long consonant and some with a short one.

The second aim of the study was to test whether exposure to orthographic input at the point of initial word learning affects metalinguistic awareness. Fewer studies had demonstrated that the presence of the orthographic form during L2 learning affects phonological awareness in comparison to the auditory input alone. For instance, Detey and Nespoulous (2008) found that L2 written forms led Japanese_{L1} speakers during syllabic segmentation to add epenthetic vowels in French_{L2} pseudowords containing consonant clusters that are not legal in L1 Japanese. This error did not occur if the pseudowords were learned without the written form (see also Young-Scholten et al., 1999). In the present study, we used a rhyme judgment task, in which participants had to decide whether a pair of lexical stimuli constituted a rhyme. The task had been previously used by Bassetti et al. (under revision) to test the effects of number of consonant letters using pairs of spoken words, and in the present study we included pseudowords. We predicted that the Phonology & Orthography group would incorrectly reject a pseudoword-word pair such as *prinnish-finish* as rhymes, because they would consider the consonant in the pseudoword as a geminate.

As the final aim, we tested whether exposure to orthographic forms increases the number of spoken words learned, compared with auditory input only. There is evidence that learning spoken words together with orthographic forms results in more word learning in a native language (Nelson et al., 2005; Miles et al., 2016). However, the evidence of this facilitative effect in a second language is limited to child beginner learners of English (Hu, 2008; Vadasy and Sanders, 2015). We predicted that the Phonology & Orthography group would learn more spoken novel words, compared to the Phonology group, and therefore show more accurate performance in the picture naming task.

MATERIALS AND METHODS

Participants

Forty-eight Italian high-school learners of English were randomly assigned to one of two groups: a Phonology group and a Phonology & Orthography group (both $n = 24$; two additional participants were eliminated because they failed to complete the learning phase described below). Both groups learned the same English spoken words and pseudowords from auditory input during the experiment, but the Phonology & Orthography group also received orthographic input.

Participants were native speakers of the Roman variety of Standard Italian, who were attending the third ($n = 41$) or fourth ($n = 7$, of which 4 in the Phonology group) year in one of two state-run high schools in Rome, a classical and a scientific high school. None of the participants knew another language with contrastive consonant length. One student reported being dyslexic, but performed similarly to the others. Participation in the study was voluntary and was rewarded with book vouchers.

Participants were studying English as a compulsory school subject for 3 h a week, using British English textbooks. They had been studying English at school for 10.2 years (SD = 1.3; Phonology group: M = 10.4, SD = 1.4; Phonology & Orthography group: M = 10.0, SD = 1.2). About half (56%) of participants had studied English in extra-scholastic settings with native teachers (Phonology group: n = 15; Med = 10 months, range = 4–64; Phonology & Orthography group: n = 12; Med = 16.5 months, range = 1–72. Kruskal-Wallis test: $\chi^2(1) = 0.54$, $p = 0.46$). About half (54%) had never been in an English-speaking country, the others reported a few weeks of study abroad (Phonology group: n = 11, Med = 2.7 weeks, range = 2–10; Phonology & Orthography group: n = 11, Med = 2 weeks, range = 2–4. Kruskal-Wallis test: $\chi^2(1) = 0.01$, $p = 0.90$).

Participants reported spending much more time listening to English than reading it (Phonology group: Listening—Med = 3 h per week, range = 0–28; Reading—Med = 1, range = 0–10. Phonology & Orthography group: Listening—Med = 4.5, range = 0–30; Reading—Med = 1, range = 0–30). They spoke English for < 1 h a week (Phonology group: Med = 0, range = 0–10; Phonology & Orthography group: Reading—Med = 0.7, range = 0–6). All but one of the participants considered a native-like English pronunciation important, very important, or extremely important. Overall, they preferred a British English accent to an American English one.

Stimuli

Materials consisted of 20 English words and 20 pseudowords (see **Supplementary Table 1**). All lexical items were monomorphemic and disyllabic, and contained a target consonant—[p], [t], or [n]—in post-tonic intervocalic position (see Bassetti, 2017; Bassetti et al., under revision). Real words were frequent lexical items for participants, as confirmed by their teachers. Pseudowords were created by changing the initial consonants of their paired word, for instance creating *prinnish* from *Finnish*. For each pseudoword C- and CC- versions were created, for instance *prinnish* and *prinish*. Half of the items were spelled with double consonants (CC) and half with a single consonant (C). There were therefore ten each of CC-words, CC-pseudowords, C-words, and C-pseudowords. Lexical items were nouns for tools, animals or plants (real words also included three adjectives), and were associated with an image of their referent.

These 40 words were used to create 30 semi-minimal pairs, each consisting of a word and a pseudoword that contained the same target consonant in the same VCV rhyme. Target consonant spelling was manipulated to obtain three types of word-pseudoword pair as follows. In CC_{pw}-C_w *pairs*, the target consonant was spelled with double letters in the pseudoword and a single consonant letter in the word, e.g., /ˈprɪnɪʃ/-/ˈfɪnɪʃ/ *(prinnish-finish)*. In C_{pw}-C_w *pairs*, the target consonant was spelled with a single consonant letter in both word and pseudoword, e.g., /ˈprɪnɪʃ/-/ˈfɪnɪʃ/ *(prinish-finish)*. In CC_w-C_{pw} *pairs*, the target consonant was spelled with double letters in the word and a single consonant letter in the pseudoword, e.g., /ˈfɪnɪʃ/- /ˈprɪnɪʃ/ *(Finnish-prinish)*. Creating pseudoword-word pairs was more efficient than creating pairs of pseudowords, because it allowed us to test our hypothesis with half as many pseudowords as would be needed to use pseudoword-pseudoword pairs. This was important, considering that learning new words in this type of experimental setting is very time-consuming (Escudero et al., 2008). By using the same pseudowords in CC_w-C_{pw} and C_{pw}-C_w pairs, it was possible to use the 40 lexical items to obtain ten each of CC_{pw}-C_w, CC_w-C_{pw}, and C_{pw}-C_w pairs.

To reduce the risk of fatigue and frustration, each participant only learned 10 pseudowords (five C- and five CC-pseudowords). They also learned the association between a picture and 15 words (ten C- and five CC-words). To achieve this, each participant learned one of two lists, each one containing five each of CC_{pw}-C_w, CC_w-C_{pw} pairs, and C_{pw}-C_w pairs. Each list contained only the C- or the CC- version of each pseudoword. For instance, *prinnish* appeared in the first list within the CC_{pw}-C_w pair *prinnish-finish*, and *prinish* in the second list within the C_{pw}-C_w pair *prinish-finish*. The list containing the C-pseudoword also contained the corresponding CC word within a CC-C pair (for instance, the list with *prinish-finish* also contained *Finnish-prinish*). The two lists were counterbalanced between participants and groups.

Lexical items were recorded by a female Southern British English native speaker and English language teacher, who read aloud words and pseudowords spelled with a singleton consonant. She was instructed to speak clearly and slightly slowly, and to pronounce pseudoword rhymes as in the corresponding word. Recording took place in a sound-attenuated room using a Røde NT2-A microphone connected to an Alesis Multimix 12 Firewire mixer.

Each lexical item was paired with a color line drawing (see **Figure 1**). Pictures were used both to illustrate the meaning of the lexical item during the learning phase, and to allow us to elicit spoken production in the testing phase (see Showalter and Hayes-Harb, 2015). Words were paired with images of their referent, and pseudowords with images of unusual tools, plants or animals, whose nouns were highly infrequent in both British English and Italian. The C- and CC-versions of the same pseudoword were paired with the same image. Images were selected from the Art Explosion library (Nova Development, 2004).

Tasks and Procedure

Participants first took part in a learning session, which consisted of a learning phase and test. Following this, there were three tasks: picture naming, rhyme judgement and spelling, administered in fixed order. Participants were tested individually in a quiet room at their school, with the researcher (the first author) present to assist, and they could pause whenever they wished during and between tasks. The whole session (including learning and experimental sessions) took ∼ 1 h.

Participants were told that they would learn new words, but they were not told that the stimuli included pseudowords. The researcher notified participants of the presence of the pseudowords at the end of the experimental session once all the tasks had been completed, and gave them a list of the images used during the tasks and the associated printed real words.

Learning Phase

In the learning phase, participants learned the 25 lexical items by seeing an image of the referent and hearing the lexical item. Participants were told that they would be asked first to memorize words paired with images, and that some of the words would be unfamiliar to them. They were instructed that following the memorization task they would be asked to recall the verbal labels when the images were presented on their own. The Phonology & Orthography group also saw the lexical item's written form simultaneously with the auditory input.

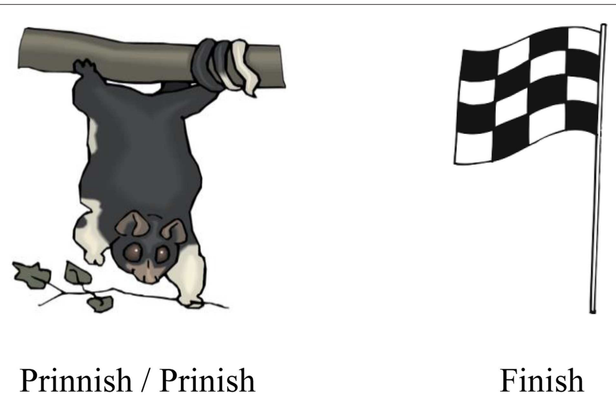


FIGURE 1 | Example of images associated to the corresponding word (right) and C- and CC-pseudowords (left).

A trial started with the presentation of a fixation point in the center of the computer screen, followed after 500 ms by an image, which remained on the screen for 3,500 ms. 1,500 ms after the onset of the image, the audio recording was played over headphones. The Phonology & Orthography group also saw the orthographic form, which appeared under the image. The onset of the orthographic form coincided with the presentation of the auditory word (i.e., 1,500 ms after the onset of the image). There was an interval of 500 ms between trials.

There were four blocks of trials; in each block stimuli were presented twice, making a total of 50 trials per block. In this way each word was presented a total of eight times across the learning phase. Trials were presented in random order within the blocks.

Learning Test

The learning test was used both to test whether the participant had learned the lexical items, and to revise unlearned items. A trial started with a fixation point in the center of the computer screen, followed after 500 ms by an image. The participant pronounced the noun aloud into a microphone. In order to check the accuracy of their response, the participant pressed the space bar. The image appeared again, and 1,000 ms after the image onset the recording was played over the headphones. The Phonology & Orthography group also saw the orthographic form, which was presented underneath the image. The image remained on the screen for a total of 2,000 ms, then the question "*Did you remember the word correctly?*" appeared on the screen, and participants answered by clicking on a tick (on the right) or a cross (on the left), both of which were presented under the written question. These answers were collected as a self-reported measure of accuracy. Each lexical item appeared four times, once in each of four blocks of trials, in random order. After each block, the screen displayed the percentage of correct answers for that block. All participants saw the four blocks, regardless of the percentage of correct answers.

Picture Naming Task

On each trial, a fixation point appeared in the center of the screen and was replaced after 500 ms by an image. The participant produced the corresponding noun aloud for three repetitions in the carrier phrase "*The word ___ should follow*," and clicked on an on-screen button after each repetition. The carrier phrase was used to keep the target word in the nuclear position within the intonational unit (Bassetti, 2017). The three repetitions were used to calculate a mean duration in order to increase reliability (Flege, 1995; Bassetti, 2017). The button presses were used to ensure that participants repeated each item three times (Bassetti, 2017). The utterances were recorded for later analysis. Each image was presented once and in random order, for a total of 25 trials.

Rhyme Judgment Task

In order to test whether participants considered the consonants spelled with double consonants as geminates, they were asked in the rhyme judgement task to judge whether two lexical items rhymed. CC_{pw}-C_w rhymes consisted of a pseudoword-word near-minimal pair, where both items had the same VCV rhyme, which was spelled with double consonants in the pseudoword and with a singleton consonant in the word, for instance /ˈprɪnɪʃ/-/ˈfɪnɪʃ/ *(prinnish-finish)*. On each trial participants saw two images side by side in the upper part of the computer screen. They were asked to recall the corresponding nouns, and decide whether they rhymed by clicking on one of two buttons below the image, a tick on the right for a rhyme, or a cross on the left for a non-rhyme. Images appeared on the left or the right interchangeably (half of the time in one of the two positions). A 500 ms black screen was used as a pause between trials. Each participant saw 20 trials: five CC_{pw}-C_w rhymes (CC pseudoword-C word), five C_{pw}-C_w rhymes (C pseudoword-C word), five CC_w-C_{pw} rhymes (CC word-C pseudoword) and five non-rhyme fillers. The fillers comprised two real words, both spelled with singleton consonant (e.g., *mini-many*). Fillers were used to add non-rhyming pairs, and to reduce the number of CC-C pairs, thereby reducing the risk of participants guessing the aim of the experiment.

Spelling Task

A spelling task was used to test whether the participants in the Phonology & Orthography group correctly remembered the spelling of the 25 lexical items, and to assess how the participants in the Phonology group spelled them. After a 500 ms fixation point, participants saw an image in the center of the screen, and typed its noun on a fixed-length line below the image. The response replaced the line. The participant pressed the return key to start the next trial. Each of the 25 images appeared once, in random order. There were no time limits, and participants were allowed to delete and retype the responses up to the beginning of the next trial. If they could not remember the word, they were allowed to skip the trial, but they were encouraged to always try to type an answer.

Equipment

All tasks were run in OpenSesame 3.1.9 Jazzy James (Mathôt et al., 2012), which managed randomization and recorded keyboard and mouse responses. Auditory input was presented over an AKG HSD171 headset. Participants' productions were recorded using a Zoom H4N Pro digital recorder connected to the headset's dynamic microphone.

Analysis

Acoustic Analysis

In order to analyze the picture naming task data, a trained phonetician measured the duration of each target sound using Praat software (Boersma and Weenink, 2016) following the standard procedure described in Bassetti et al. (2018).

Statistical Analysis

Data were analyzed using R software, version 3.4.4 (RStudio Team, 2018) with RStudio 1.1.447 (R Core Team, 2018). All the following analyses were performed using (linear or generalized) mixed effect models, with package *lmerTest* (Kuznetsova et al., 2017—*lme4* version: Bates et al., 2015b). As a first step, all the models included all the fixed effects of interest. Then, model reduction was performed through likelihood ratio test (Baayen et al., 2008). Initially, the maximal random effects structure was considered (Barr et al., 2013). In case of failure of convergence or overfitting (random effects were perfectly collinear), we proceeded with model reduction following Bates et al. (2015a). In the Results section we report only the final fixed and random effect structures. *P*-values for *t*-statistics were obtained using Satterthwaite's method for denominator degrees of freedom (provided by the lmerTest package). Conditional and marginal R^2 were calculated using function r.squaredGLMM in the *MuMIn* package (Bartoń, 2018), whereas Tukey's *post-hoc* contrasts were performed using function contrast in the *lsmeans* package (Lenth, 2016).

Learning test

A generalized linear mixed model with binomial error distribution was used to analyze the number of spoken pseudowords learned in the learning test. The dependent variable was whether the participant rated their answer as correct (coded as 1) or incorrect (coded as 0) after hearing the correct answer. The responses of five Phonology group participants were lost due to technical issues.

Picture naming task

4.9% of the data (59 responses out of a total of 1,200) were lost because the answer was missing or mispronounced, or the recording was not suitable for acoustic analysis due to background noise or interruptions. An additional 4.8% (58 responses) were removed because the response for the corresponding item in the Spelling Task was incorrect (misspelled real words, pseudowords spelled with wrong intervocalic context of the target CC-C). This removal was considered as necessary because orthographic effects on consonant production were only expected when the speaker knew the correct correspondence between the pictures and the items and the correct letters in the items (i.e., target consonants and intervocalic context, see *Spelling Task* paragraph below in this section for a more detailed rationale of error categorization in pseudowords). The final dataset contained 1,083 items. A mean duration of the target sound was calculated on the three repetitions made by the participants for each item. Data analysis was performed on consonant duration ratios which were calculated for each participant for each CC_{pw}-C_w, C_{pw}-C_w, and CC_w-C_{pw} pair. For instance, for the CC_{pw}-C_w pairs *prinnish-finish*, the ratio was calculated by dividing the duration of [n] in *prinnish* by the duration of [n] in *finish*.

Outliers were considered as ratios that were beyond the 99th percentile and below the 1st percentile (2.05% of data). The remaining ratios were log-transformed to approximate normal distribution and examined with linear mixed regression models. Model details are provided in the Results section.

Rhyme judgment task

Pairs that contained words and pseudowords which were misspelled in the Spelling Task were removed from the analysis (15.5% of data, see Picture Naming Task above and Spelling Task below for the rationale). After removing incorrect spellings and discarding the fillers pairs, 608 pairs (out of a total of 720) were included in the analysis. We coded correct answers as 1 and incorrect answers as 0. For the analysis we used a generalized mixed model with binomial error distribution.

Spelling task

A generalized linear mixed model with binomial error distribution was used to analyze the number of pseudowords spelled correctly in the Spelling Task. Reponses were coded as 1 if correct, or 0 if incorrect, and these values were used as the dependent variable in the model. We coded pseudoword spellings as incorrect if the participant had not provided an answer, or answered with a lexical item other than the one represented by the picture, or provided a spelling that was substantially different from the phonological form of the target, for instance if the target consonant was surrounded by a wrong vocalic context. With regards to the spelling of the target consonant, we treated the two participant groups differently. If the target consonant was spelled with the incorrect number of letters, we coded the answer as incorrect in the Phonology & Orthography group. For the Phonology group, we accepted both single and double consonants as correct, because the group had not learned the correct spelling and both spellings are acceptable.

RESULTS

Learning Test

The first task participants performed measured the number of pseudowords they learned. **Figure 2** shows the mean proportion of pseudowords participants in the two groups reported having correctly produced over the four learning blocks. In the generalized linear mixed model, we inserted as fixed effects Group (Phonology, Phonology & Orthography) and Block (1, 2, 3, and 4) and their interaction. As random effects we used the intercepts for subjects and items.

The interaction was not significant and was removed from the model, so the final model included the effects of Block and Group. As **Table 1** shows, the Phonology & Orthography group learned more pseudowords than the Phonology group overall, and the lack of a significant interaction shows that this effect applied across the four learning blocks. Both groups learned increasingly more pseudowords over the four learning blocks.

TABLE 1 | Results of mixed-model analysis of the effects of group (Phonology, Phonology & Orthography) and learning block (1, 2, 3, 4) on number of pseudowords learned.

Random effects		Variance	*SD*		
Participants	Intercept	2.42	1.55		
Word	Intercept	0.47	0.69		
Fixed effects		**Estimate**	***SE***	***z*-value**	***p***
Intercept		1.21	0.42	2.85	0.004**
Group (P&O)		1.21	0.52	2.31	0.020*
Block					
	(Block 2)	0.53	0.20	2.68	0.007**
	(Block 3)	1.15	0.22	5.29	<0.001***
	(Block 4)	1.59	0.24	6.70	<0.001***
		Marginal R^2:	0.10	Conditional R^2:	0.52

$^*p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

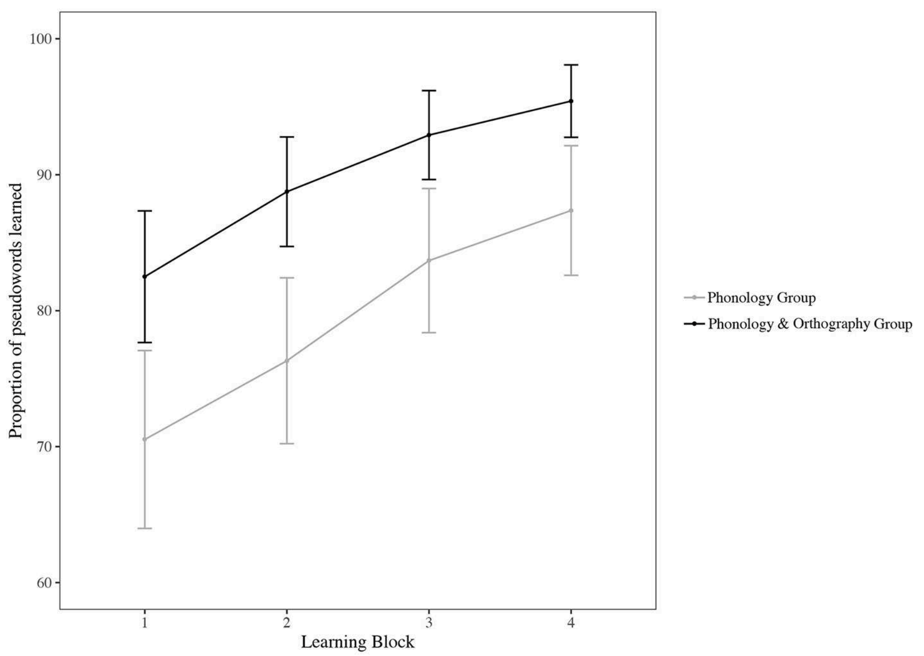


FIGURE 2 | Mean percentage of pseudowords recalled and correctly produced as a function of group (Phonology, Phonology & Orthography) and learning block (1, 2, 3, 4). Bars represent 95% *CI*s.

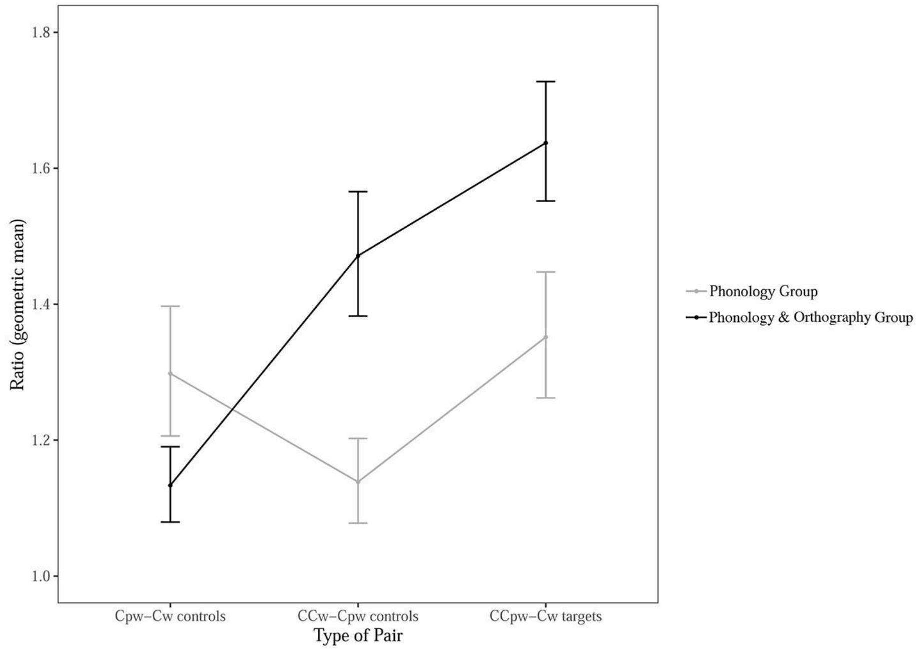


FIGURE 3 | The geometric mean of consonant duration ratio as a function of group (Phonology, Phonology & Orthography) and type of pair (C_{pw}-C_w, CC_w-C_{pw}, CC_{pw}-C_w). Bars represent 95% *CI*s.

Picture Naming Task

This was the critical task because it tested whether orthographic input affected the phonological form of the item that was learned. **Figure 3** shows the results. In line with predictions, for the Phonology & Orthography group, the geometric mean of consonant duration ratios in CC_{pw}-C_w pairs was 1.64 (*CI* [1.55–1.73]), while the mean ratio of the C_{pw}-C_w pairs was close to one ($M = 1.13$, *CI* [1.08–1.19]), and the mean ratio of the CC_w-C_{pw} pairs was close to the mean ratio of the CC_{pw}-C_w pairs ($M = 1.47$, *CI* [1.38–1.57]). For the Phonology group the geometric mean of the consonant duration ratios was 1.35 in CC_{pw}-C_w pairs (*CI* [1.26–1.45]), 1.30 in C_{pw}-C_w pairs (*CI* [1.21–1.40]), and 1.14 in CC_w-C_{pw} pairs (*CI* [1.08–1.20]).

The linear mixed model included as fixed effects Group (Phonology, Phonology & Orthography), Type of Pair (CC_{pw}-C_w, C_{pw}-C_w vs. CC_w-C_{pw} pair) and their interaction. As random effect we used the intercepts for subjects and items and the by-subject random slope for Type of Pair. The results are summarized in **Table 2**.

In confirmation of our hypothesis, multiple contrasts within groups with Tukey adjustment revealed that in the Phonology & Orthography group CC_{pw}-C_w pairs had a higher ratio than C_{pw}-C_w pairs ($t = 5.54$, $p < 0.001$) and a similar ratio as CC_w-C_{pw} pairs ($p > 0.05$). C_{pw}-C_w and CC_w-C_{pw} pairs differed significantly ($t = -3.37$, $p = 0.004$). These results revealed that the participants in the Phonology & Orthography group produced longer consonant durations for CC words and pseudowords than the consonant durations for the paired C words and pseudowords. For the Phonology group there were no statistical differences for CC_{pw}-C_w and C_{pw}-C_w, CC_{pw}-C_w and CC_w-C_{pw}, C_{pw}-C_w and CC_w-C_{pw} pairs.

Multiple comparisons between groups showed that the Phonology & Orthography group produced a longer ratio for

TABLE 2 | Results of mixed-model analysis of the effects of group (Phonology, Phonology & Orthography) and type of pair (CC_{pw}-C_w, C_{pw}-C_w, CC_w-C_{pw}) on consonant duration ratio.

Random effects		**Variance**	***SD***	***Corr***	
Participants	Intercept	0.01	1.10		
	Type of pair				
	($C_{pw\text{-}w}$)	0.02	0.13	−0.31	
	(CC_w-C_{pw})	0.01	0.10	−0.27	−0.83
Word	Intercept	0.01	0.11		
Residual		0.06	0.25		
Fixed effects		**Estimate**	***SE***	***t*-value**	***p***
Intercept		0.28	0.05	5.87	<0.001***
Group (P&O)		0.21	0.05	4.43	<0.001***
TRIAL CONDITION					
	(C_{pw}-C_w)	−0.39	0.07	−0.57	ns
	(CC_w-C_{pw})	−0.15	0.07	−2.32	0.025*
GROUP*TRIAL CONDITION					
	P&O*C_{pw}-C_w	−0.32	0.06	−4.94	<0.001***
	P&O*CC_w-C_{pw}	0.04	0.06	0.75	ns
		Marginal R^2:	0.17	Conditional R^2:	0.42

**p* < 0.05; ****p* < 0.001.

both CC_{pw}-C_w ($t = 4.43$, $p < 0.001$) and CC_w-C_{pw} ($t = 4.91$, $p < 0.001$) compared with the Phonology group. C_{pw}-C_w pair ratios were lower for the Phonology & Orthography group than the Phonology group ($t = -2.04$, $p = 0.049$).

Rhyme Judgment Task

In the Phonology & Orthography group the mean percentage of correct responses was 69% for CC_{pw}-C_w pairs ($SD = 46$, $CI = 8$), 93% for C_{pw}-C_w pairs ($SD = 26$, $CI = 5$), and 64% for CC_w-C_{pw} pairs ($SD = 48$, $CI = 9$). In the Phonology group the mean percentage was 69% for CC_{pw}-C_w rhymes ($SD = 47$, $CI = 9$), 77% for C_{pw}-C_w rhymes ($SD = 42$, $CI = 9$) and 87% for CC_w-C_{pw} rhymes ($SD = 0.33$, $CI = 0.7$). **Figure 4** summarizes the results.

In the generalized linear mixed model, we inserted as fixed effects Group (Phonology, Phonology & Orthography) and Type of Pair (CC_{pw}-C_w, C_{pw}-C_w, CC_w-C_{pw} pair) and their interaction. As random effects we used the intercept for subjects and items (words and pseudowords) and the by-subject random slope for Type of Pair. The results are summarized in **Table 3**.

Multiple contrasts within groups with Tukey adjustment revealed that the Phonology & Orthography group was more accurate with C_{pw}-C_w than CC_{pw}-C_w rhymes ($z = 3.42$, $p = 0.002$), and CC_w-C_{pw} accuracy did not differ from CC_{pw}-C_w accuracy ($p > 0.05$). C_{pw}-C_w and CC_w-C_{pw} rhymes differed significantly ($z = -3.98$, $p < 0.001$). Participants in the Phonology & Orthography group erroneously rejected CC-C rhymes (both CC_{pw}-C_w and CC_w-C_{pw}) more often than C-C rhymes, presumably because they considered consonants spelt with a single letter and those spelt with double letters as different phonemes.

In the Phonology group accuracy was lower with CC_{pw}-C_w than CC_w-C_{pw} pairs ($z = -2.44$, $p = 0.04$). There were no differences in accuracy between CC_{pw}-C_w and C_{pw}-C_w pairs and between CC_w-C_{pw} and C_{pw}-C_w pairs ($p > 0.05$).

Multiple contrasts between groups showed that accuracy did not differ between groups for CC_{pw}-C_w pairs ($p > 0.05$), but the Phonology & Orthography group was more accurate with C_{pw}-C_w pairs ($z = 2.79$, $p = 0.005$) and less accurate with CC_w-C_{pw} pairs ($z = -3.33$, $p < 0.001$) compared with the Phonology group.

Spelling Task

In the spelling task, the average number of pseudowords spelled correctly was higher in the Phonology & Orthography group ($M = 97\%$, $SD = 18$) than in the Phonology group ($M = 86\%$, $SD = 35$). In the generalized linear mixed model we inserted Group (Phonology, Phonology & Orthography) as fixed effect, and the intercept for participants and pseudowords as random effects. The final model, shown in **Table 4**, reveals that the Phonology & Orthography group spelled more pseudowords correctly than the Phonology group.

In order to understand whether participants perceive English consonants as short or long in words whose spelling they do not know, we analyzed the correct pseudoword spellings produced by the Phonology group. The Phonology group spelled just over half of pseudowords with double consonants (59%, or 122 out of 206 valid pseudoword spellings). A Poisson generalized linear model on count data as the dependent variable and Spelled Consonant (CC vs. C) as fixed effect revealed that the Phonology group spelled pseudowords with double consonants

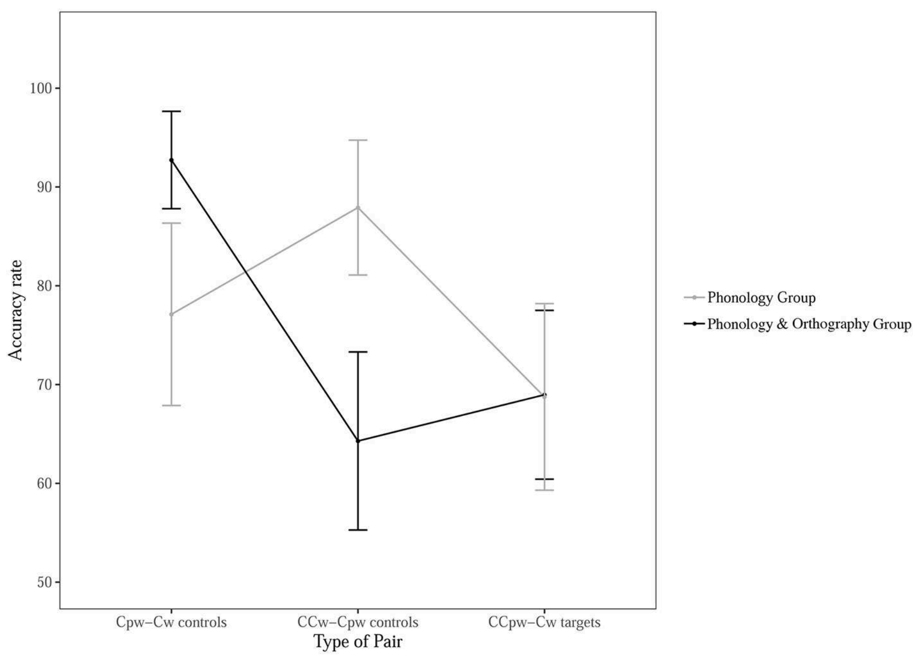


FIGURE 4 | The proportion of correct responses as a function of group (Phonology, Phonology & Orthography) and type of pair (C_{pw}-C_W, CC_W-C_{pw}, CC_{pw}-C_W) in the rhyme judgement task. Bars represent 95% *CI*s.

TABLE 3 | Results of mixed-model analysis of the effects of group (Phonology, Phonology & Orthography) and type of pair (CC_{pw}-C_w, C_{pw}-C_w, CC_w-C_{pw}) on mean percentage of correct responses in the rhyme judgment task.

Random effects		**Variance**	***SD***	***Corr***	
Participants	Intercept	0.69	0.83		
	Type of pair				
	(C_{pw}-C_w)	1.24	0.11	−0.80	
	(CC_w-C_{pw})	0.58	0.76	0.33	0.30
Word	Intercept	0.65	0.81		
Fixed effects		**Estimate**	***SE***	***z*-value**	***p***
Intercept		0.99	0.41	2.42	0.015**
Group (P&O)		−0.10	0.41	−0.25	ns
TRIAL CONDITION					
	(C_{pw}-C_w)	0.49	0.59	0.83	ns
	(CC_w-C_{pw})	1.60	0.64	2.51	0.012*
GROUP*TRIAL CONDITION					
	P&O*C_{pw}-C_w	1.40	0.70	2.00	0.046*
	P&O*CC_w-C_{pw}	−1.83	0.61	−2.96	0.003**
		Marginal R^2:	0.13	Conditional R^2:	0.41

*$p < 0.05$; **$p < 0.01$.

TABLE 4 | Results of mixed-model analysis of the effects of group (Phonology, Phonology & Orthography) on number of pseudowords correctly spelled.

Random effects		**Variance**	***SD***		
Participants	Intercept	5.01	2.24		
Word	Intercept	1.89	1.38		
Fixed effects		**Estimate**	***SE***	***z*-value**	***p***
Intercept		3.42	0.81	4.24	<0.001***
Group (P&O)		2.65	0.97	2.73	0.006**
		Marginal R^2:	0.00	Conditional R^2:	0.06

**$p < 0.01$; ***$p < 0.001$.

39% more often than with a single consonant (Estimate = 0.33, $z = 2.33$, $p = 0.020$).

Further Analyses

The fact that the Phonology group spelled some pseudowords with double consonants and some with a singleton consonant led us to hypothesize that this group learned some pseudowords as containing long consonants and others as containing short consonants. If this is true, then these participants should produce longer consonants in pseudowords they spelled with double letters than in pseudowords they spelled with singleton consonant, and consider the former as geminates and the latter as singleton consonants. To test this hypothesis, we categorized each CC_{pw}-C_w and C_{pw}-C_w pair produced by each Phonology group participant according to how s/he had spelled the pair in the Spelling Task (see e.g. Soković-Perović et al., 2019). If the participant spelled the pseudoword with double letters, we classified the pair as a CC-C pair, and if the participant spelled the pseudoword with a single consonant we classified the pair as a C-C pair. Therefore, we predicted that in both production and awareness the Phonology group's CC-C pairs should behave similarly to the Phonology & Orthography's group CC_{pw}-C_w pairs, and the former's C-C pairs should behave similarly to the latter's C_{pw}-C_w pairs.

In the Picture Naming Task, the Phonology group's geometric mean ratios were 1.08 for C-C pairs (*CI* [1.02–1.14]) and 1.56 for CC-C pairs (*CI* [1.47–1.65]). These ratios were very similar to the Phonology & Orthography group's C_{pw}-C_w pairs and CC_{pw}-C_w pairs. We then ran a linear mixed model, inserting Group (Phonology, Phonology & Orthography), Type of Pair (C-C, CC-C), and their interaction as fixed effects, and the intercepts for subjects and items as random effects. Both Group and the interaction were removed from the model because they did not improve model fit. The effect of Type of Pair (Estimate = 0.35, $SE = 0.03$, $t = 11.55$, $p < 0.001$) showed that C-C pairs were produced with smaller ratios than CC-C for both groups, confirming that participants produced longer consonants when they spelled the pseudoword with double letters.

In the Rhyme Judgment Task, the Phonology group's mean accuracy was higher (81%) for C-C pairs ($SD = 40$, $CI = 9$) than for CC-C pairs ($M = 66\%$, $SD = 47$, $CI = 9$). These figures are very similar to the Phonology & Orthography group's accuracy levels for C_{pw}-C_w and CC_{pw}-C_w pairs, respectively. We then ran a generalized linear mixed model with Group (Phonology, Phonology & Orthography), Type of Pair (C-C, CC-C), and their interaction as fixed effects, and the intercepts for subjects and items as random effects. Both Group and the interaction were removed from the model because they did not improve model fit. The effect of Type of Pair (Estimate = −1.24, $SE = 0.48$, $z = -2.60$, $p = 0.009$) shows that both groups were more accurate with C-C than with CC-C pairs.

DISCUSSION

Orthographic forms affect the perception and production of spoken words in second language learners, but it is not clear whether these effects can be established at the point of first learning the word. The main goal of the current study was then to investigate how the simultaneous presentation of orthographic and phonological inputs affects the learning of novel $English_{L2}$ words, and, in particular, we tested whether the presence of a double consonant in the spelling of a new $English_{L2}$ word may lead $Italian_{L1}$ speakers to perceive and subsequently produce this word with a longer consonant than a word spelled with a singleton consonant, therefore producing a contrast that does not exist in the L2 auditory input or in the L2 phonological system. Results revealed that this was indeed the case.

The first aim of the study was to test whether learning a new word that is spelled with double consonant letters results in producing the new word with a longer consonant than a similar word that is spelled with a singleton consonant letter.

The Phonology & Orthography group, who had learned the novel words' spoken and written form, produced longer consonants in such words, compared to the Phonology group who had learned only the spoken form. This was shown by the interaction between group and type of word pair, and by multiple comparisons between groups. CC_{pw}-C_w and CC_w-C_{pw} pairs had a higher consonant duration ratio in the Phonology & Orthography group than in the Phonology group. In addition, the Phonology & Orthography group produced novel CC-words with longer consonants and novel C-words with shorter consonants. This was shown by the high ratios of CC_{pw}-C_w and CC_w-C_{pw} pairs (respectively 1.64 and 1.47), while the ratio in C_{pw}-C_w pairs was just above one (1.13). The high consonant ratios in pairs containing a double consonant show that the double consonant was produced as a geminate, in both real words and newly-learned pseudowords. This confirms previous findings that $Italian_{L1}$ speakers of $English_{L2}$ produce known English words with a geminate consonant when that consonant is spelled with double letters (Bassetti, 2017; Bassetti et al., 2018), and crucially shows that such effects are found in newly-learned words. The presence of orthographic input during word learning presumably led to the activation of L1 phoneme-grapheme correspondence rules and their transposition to the newly learned L2 words, confirming a strong influence of orthography on L2 production when L1-L2 incongruent graphemes are presented (Hayes-Harb et al., 2010; Pytlyk, 2011; Escudero et al., 2014; Showalter and Hayes-Harb, 2015; Mathieu, 2016).

The Phonology group, who had not seen the items' orthographic forms, produced the novel words' target consonant with similar duration across conditions, and indeed multiple contrasts within group revealed that the mean ratios did not differ across types of pairs. However, crucially, in comparison with the Phonology & Orthography group, the Phonology group produced consonants with a smaller duration ratio in CC_{pw}-C_w (1.35 vs. 1.64) and CC_w-C_{pw} pairs (1.14 vs. 1.47), and with a higher ratio in C_{pw}-C_w pairs (1.30 vs. 1.13). These results support Bassetti's (2017) hypothesis that, in the absence of orthographic input, Italians categorize the duration of English consonants in native speaker's production as in-between short and long consonants, as follows. The Phonology group had a smaller ratio than the Phonology & Orthography group in CC_{pw}-C_w pairs, because both groups produced C-words with a singleton consonant, but the Phonology & Orthography group generally produced CC-pseudowords with a geminate, and the Phonology group generally produced about half CC-pseudowords with a geminate and half with a singleton. Similarly, CC_w-C_{pw} pairs had smaller ratios in the Phonology than in the Phonology & Orthography group because both groups produced real CC-words with a geminate, but the Phonology & Orthography group produced all C-pseudowords with a singleton, whereas the Phonology group produced many of these as a geminate. Finally, with C_{pw}-C_w pairs the Phonology & Orthography group had a smaller ratio than the Phonology group because the former produced all C-pseudowords with a singleton, whereas the latter produced about half of the C-pseudowords with a geminate. Further evidence comes from the re-analysis of the data from the picture naming task, where the Phonology group's stimulus pairs were categorized according to the participant's spelling as CC-C and C-C pairs, rather than based on the pseudoword's spelling we had taught to the Phonology & Orthography group. When we compared the Phonology group's ratios with pairs they had spelled as CC-C and C-C, they performed similarly to the Phonology & Orthography group's performance with CC_{pw}-C_w and C_{pw}-C_w pairs. This finding demonstrated that, in the absence of orthographic input, Italian speakers can perceive English consonants as either singleton or geminate consonants, and this was reflected in both their spoken and written production.

The second aim of the study was to test whether the same orthographic effect on new word learning found in speech production would also be found in metalinguistic awareness. Indeed, the results of the rhyme judgment task were in line with the results from the production task. As predicted, the Phonology & Orthography group incorrectly rejected CC_{pw}-C_w rhymes more often than C_{pw}-C_w rhymes, and as often as CC_w-C_{pw} rhymes, showing that they correctly accepted rhymes where the target consonant is spelled with a singleton letter (C_{pw}-C_w rhymes, average accuracy of 93%) and incorrectly rejected rhymes where the target consonant is spelled with a singleton letter in one item and double letters in the other one (CC_{pw}-C_w and CC_w-C_{pw} rhymes, 69% and 64% correct, respectively). This is presumably because the Phonology & Orthography group erroneously interpreted the presence of a double consonant in word and pseudoword spellings as a long sound and rejected rhymes containing the same consonant spelled with singleton or double consonant, because long and short consonants are different phonemes in their native language. This confirms findings by Bassetti et al. (under revision) with real words, and extends such findings to newly-learned words.

The Phonology group were instead most accurate with CC_w-C_{pw} rhymes, but showed no difference in accuracy between CC_{pw}-C_w and C_{pw}-C_w rhymes. This is because this group—as shown in the production task—interpreted more than half of novel words as containing a geminate consonant. Therefore, CC_{pw}-C_w and C_{pw}-C_w rhymes both contained a real C-word and a novel word that was sometimes evaluated to contain a geminate and sometimes a singleton. CC_w-C_{pw} rhymes were most often accepted because to this group all CC-words contained a geminate and more than half of the pseudowords also contained a geminate. We performed further analysis of the Phonology group's rhyme judgments based on how each participant spelled the pseudowords in the spelling task, as we had done with the data from picture naming, and found that for pairs containing a single-consonant real word the Phonology group rejected more pairs when they had spelled the pseudoword with double consonant letters than when they had spelled the pseudoword with a singleton consonant. This is similar to what the Phonology & Orthography group did with CC_{pw}-C_w vs. C_{pw}-C_w rhymes.

It can be concluded that orthographic input affects L2 word acquisition, not only in speech production, but also in a similar way in metalinguistics awareness. This is an interesting finding because there has been very little research on orthographic effects on metalinguistic awareness in second language speakers.

As a third aim, we tested whether the presence of orthographic forms during learning would result in learning more L2 words, compared with auditory input only. Results from the learning and spelling tests show that the Phonology & Orthography group learned more novel words (pseudowords) than the Phonology group. Perhaps more predictably, those who had seen the words' written forms could spell more words than those who had only had auditory input and were guessing the words' spelling. Crucially, the Phonology & Orthography group also learned more spoken words than the Phonology group. This was evident at all four time points during the learning phase. The Phonology & Orthography group learned on average more than 90% of the novel words after just one exposure—compared with about 85% in the Phonology group – and reached ceiling level after just four exposures—compared with around 90% accuracy in the Phonology group. While both groups showed progression in learning, with more words correctly produced in the fourth than in the first repetition, the Phonology group's ultimate performance was similar to the Phonology & Orthography group's performance after first exposure. While previous research found that spoken L2 vocabulary acquisition is more efficient with than without orthographic input in child beginner learners of English (Hu, 2008; Vadasy and Sanders, 2015), the present study shows this facilitative effect in adult experienced L2 learners.

The results of the learning test cannot be considered conclusive, because the analysis is based on self-reported data, whereby participants produced the word from memory, heard the native speaker's model, and evaluated whether they had produced the correct form or not. Future research could use more objective measures of word learning. However, overall our results show that orthographic input results in almost 100% accuracy in spoken and written word learning after just four exposures.

CONCLUSIONS

This study showed for the first time that the presence of orthographic input during the initial learning of a second language spoken word can lead experienced L2 speakers to learn the phonological form of the word with a sound that does not exist in the auditory input they were exposed to, or indeed in the target language. It is possible that orthographic input results in a perceptual illusion, such that L2 speakers perceive—and therefore learn—a novel word as containing a long consonant if it is spelled with double letters. This is due to recoding the L2 orthographic word using L1 grapheme-phoneme conversion (what Hayes-Harb and others call "orthographic incongruency"). It appears that the effects of orthographic forms on L2 word production that have been widely reported are established in the very early stages of L2 word learning.

While it is perhaps to be expected that those without previous experience of a language's phonology and orthography would fall back on L1 phonology and orthography to make sense of L2 input, we found such effects in learners with over 10 years' exposure to the second language. Furthermore, orthographic effects have been shown in experienced L2 speakers learning novel words, but these effects resulted in producing or perceiving an incorrect L2 phonological category, whereas here the orthographic effect resulted in the production of a sound that does not exist in the target language. Finally, previous research on orthographic effects on L2 speech production could not explain whether the effect of double letters on consonant length in the production of L2 words was due to repeated exposure or whether it was established at the point of first learning the word. Results from the present study indicated that just four presentations were sufficient to establish a phonological representation containing a sound not present in the auditory input.

We found effects of the same orthographic form (double letters) on speech production, metalinguistic awareness and spelling of newly learned words. These findings confirm Bassetti et al.'s (under revision) findings with real words, and support Bassetti's (2006, 2008) view that L2 phonological representations are affected by both L2 orthographic input, reinterpreted according to L1 orthography-phonology correspondence, and phonological input, reinterpreted according to the L1 phonological system. In this case, learners apply the L1 correspondence between double letters and long consonants, as well as the L1 distinction between long and short consonants, to the English language, where the distinction and correspondence do not exist. The presence of the same orthographic effect on spoken production, written production and metalinguistic awareness is powerful evidence that orthography affects L2 phonological representations, even in newly learned words.

ETHICS STATEMENT

The study was carried out in accordance with the recommendations of the British Psychological Society. Prior to data collection ethical approval was obtained from the Humanities & Social Sciences Research Ethics Committee of the University of Warwick (UK). All participants signed a written informed consent form.

AUTHOR CONTRIBUTIONS

TC and BB contributed to the design of the study and the interpretation of results. TC was responsible for data collection and analysis and drafted the manuscript. BB and JM revised it critically for important intellectual content. All the authors agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

FUNDING

This study was supported by a Leverhulme Trust Research Grant [grant RPG 2013 180] awarded to BB and JM.

ACKNOWLEDGMENTS

The authors are grateful to Paolo Mairano and Rosalba Nodari for performing the acoustic analysis, to Jo Gakonga for recording auditory stimuli, and to all participants and their teachers.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2019.00031/full#supplementary-material

REFERENCES

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Memory Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Memory Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bartoń, K. (2018). *MuMIn: Multi-Model Inference. Package version 1.42.1.* Available online at: https://CRAN.R-project.org/package=MuMIn

Bassetti, B. (2006). Orthographic input and phonological representations in learners of Chinese as a foreign language. *Written Lang Liter.* 9, 95–114. doi: 10.1075/wll.9.1.07bas

Bassetti, B. (2007). "Effects of hanyu pinyin on pronunciation in learners of Chinese as a Foreign Language," in *The Cognition, Learning and Teaching of Chinese Characters,* eds A. Guder, X. Jiang and Y. Wan (Beijing: Beijing Language and Culture University Press), 156–179.

Bassetti, B. (2008) "Orthographic input and second language phonology," in *Input Matters in SLA* eds T. Piske and M. YoungScholten (Clevedon: Multilingual Matters), 191–206. doi: 10.21832/9781847691118-013

Bassetti, B. (2017). Orthography affects second language speech: double letters and geminate production in English. *J. Exp. Psychol. Learn. Memory Cogn.* 43, 1835–1842. doi: 10.1037/xlm0000417

Bassetti, B., and Atkinson, N. (2015). Effects of orthographic forms on pronunciation in experienced instructed second language learners. *Appl. Psycholinguistics* 36, 67–91. doi: 10.1017/S0142716414000435

Bassetti, B., Soković-Perović, M., Mairano, P., and Cerni, T. (2018). Orthography-induced length contrasts in the second language phonological systems of L2 speakers of English: Evidence from minimal pairs. *Lang. Speech* 61, 577–597. doi: 10.1177/0023830918780141

Bates, D., Kliegl, R., Vasishth, S., and Baayen, H. (2015a). *Parsimonious Mixed Models*. arXiv preprint arXiv:1506.04967.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015b). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Boersma, P., and Weenink, D. (2016). *Praat: Doing Phonetics by Computer [Computer program]. Version 6.0.19.* Available online at: http://www.praat.org/

Chambré, S. J., Ehri, L. C., and Ness, M. (2017). Orthographic facilitation of first graders' vocabulary learning: does directing attention to print enhance the effect? *Read. Writ.* 30, 1137–1156. doi: 10.1007/s11145-016-9715-z

Detey, S., and Nespoulous, J. (2008). Can orthography influence second language syllabic segmentation? Japanese epenthetic vowels and French consonantal clusters. *Lingua* 118, 66–81. doi: 10.1016/j.lingua.2007.04.003

Ehri, L. C., and Wilce, L. S. (1979). The mnemonic value of orthography among beginning readers. *J. Educ. Psychol.* 71, 26–40.doi: 10.1037/0022-0663.71.1.26

Erdener, V. D., and Burnham, D. K. (2005). The role of audiovisual speech and orthographic information in nonnative speech production. *Lang. Learn.* 55, 191–228. doi: 10.1111/j.0023-8333.2005.00303.x

Escudero, P., Hayes-Harb, R., and Mitterer, H. (2008). Novel second-language words and asymmetric lexical access. *J. Phonet.* 36, 345–360. doi: 10.1016/j.wocn.2007.11.002

Escudero, P., Simon, E., and Mulak, K. E. (2014). Learning words in a new language: orthography doesn't always help. *Bilingualism Lang. Cogn.* 17, 384–395. doi: 10.1017/S1366728913000436

Flege, J. E. (1995). "Second language speech learning: theory, findings and problems," in *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues,* ed W. Strange (Timonium, MD: York Press), 233–277.

Hayes-Harb, R., Nicol, J., and Barker, J. (2010). Learning the phonological forms of new words: effects of orthographic and auditory input. *Lang. Speech* 53, 367–381. doi: 10.1177/0023830910371460

Hu, C.-F. (2008). Use orthography in L2 auditory word learning: who benefits? *Read. Writing* 21, 823–841. doi: 10.1007/s11145-007-9094-6

Kuznetsova, A., Brockhoff,. P.B., and Christensen, R.H.B. (2017). lmerTest Package: tests in linear mixed effects models. *J. Stat. Soft* 82, 1–26. doi: 10.18637/jss.v082.i13

Lenth, R. V. (2016). Least-squares means: the R package lsmeans. *J. Stat. Softw.* 69, 1–33. doi: 10.18637/jss.v069.i01

Mathôt, S., Schreij, D., and Theeuwes, J. (2012). OpenSesame: an open-source, graphical experiment builder for the social sciences. *Behav. Res. Methods* 44, 314–324. doi: 10.3758/s13428-011-0168-7.

Mathieu, L. (2016). The influence of foreign scripts on the acquisition of a second language phonological contrast. *Second Lang. Res.* 32, 145–170. doi: 10.1177/0267658315601882

Miles, K. P., Ehri, L. C., and Lauterbach, M. D. (2016). Mnemonic value of orthography for vocabulary learning in monolinguals and language minority English-speaking college students. *J. College Read. Learn.* 46, 99–112. doi: 10.1080/10790195.2015.1125818

Nelson, J. R., Balass, M., and Perfetti, C. A. (2005). Differences between written and spoken input in learning new words. *Written Lang. Literacy* 8, 25–44. doi: 10.1075/wll.8.2.04nel

Nova Development (2004). *Art Explosion Library*. Calabasas, CA: Nova Development Co.

Pytlyk, C. (2011). Shared orthography: do shared written symbols influence the perception of L2 sounds? *Modern Lang. J.* 54, 541–557. doi: 10.2307/41413379

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org/

Ricketts, J., Bishop, D. V., and Nation, K. (2009). Orthographic facilitation in oral vocabulary acquisition. *Q. J. Exp. Psychol.* 62, 1948–1966. doi: 10.1080/17470210802696104

Rosenthal, J., and Ehri, L. C. (2008). The mnemonic value of orthography for vocabulary learning. *J. Educ. Psychol.* 100:175. doi: 10.1016/j.jecp.2015.01.015

RStudio Team (2018). *RStudio: Integrated Development for R*. Boston, MA: RStudio, Inc. Available online at: http://www.rstudio.com/

Showalter, C., E., and Hayes-Harb, R. (2013). Unfamiliar orthographic information and second language word learning: a novel lexicon study. Second *Lang. Res.* 29, 185–200. doi: 10.1177/0267658313480154

Showalter, C. E., and Hayes-Harb, R. (2015). Native English speakers learning Arabic: the influence of novel orthographic information on second language phonological acquisition. *Appl. Psycholinguistics* 36, 23–42. doi: 10.1017/S0142716414000411

Soković-Perović, M., Bassetti, B., and Dillon, S. (2019). English orthographic forms affect L2 English speech production in native users of a non-alphabetic writing system. *Bilingual. Lang. Cogn.* 1–11. doi: 10.1017/S136672891900035X

Vadasy, P. F., and Sanders, E. A. (2015). Incremental learning of difficult words in story contexts: the role of spelling and pronouncing new vocabulary. *Read. Writing* 28, 371–394. doi: 10.1007/s11145-014-959-9

Vokic, G. (2011). When alphabets collide: alphabetic first-language speakers' approach to speech production in an alphabetic second language. Second *Lang. Res.* 27, 391–417. doi: 10.1177/0267658310396627

Young-Scholten, M., Akita, M., and Cross, N. (1999). "Focus on form in phonology: orthographic exposure as a promoter of epenthesis," in *Pragmatics and pedagogy. Proceedings of the third PacSLRF,* eds P. Robinson andamp and N. O. Jungheim (Tokyo: Aoyama Gakuin University), 2, 227–233.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Communication

ORIGINAL RESEARCH
published: 10 September 2019
doi: 10.3389/fcomm.2019.00048

A Processing-Oriented Investigation of Inflectional Complexity

Claudia Marzi, Marcello Ferro* and Vito Pirrelli*

Institute for Computational Linguistics, National Research Council, Pisa, Italy

Due to the typological diversity of their inflectional processes, some languages are intuitively more difficult than other languages. Yet, finding a single measure to quantitatively assess the comparative complexity of an inflectional system proves an exceedingly difficult endeavor. In this paper we propose to investigate the issue from a processing-oriented standpoint, using data processed by a type of recurrent neural network to quantitatively model the dynamic of word processing and learning in different input conditions. We evaluate the relative complexity of a set of typologically different inflectional systems (Greek, Italian, Spanish, German, English and Standard Modern Arabic) by training a Temporal Self-Organizing Map (TSOM), a recurrent variant of Kohonen's Self-Organizing Maps, on a fixed set of verb forms from top-frequency verb paradigms, with no information about the morphosemantic and morphosyntactic content conveyed by the forms. After training, the behavior of each language-specific TSOM is assessed on different tasks, looking at self-organizing patterns of temporal connectivity and functional responses. Our simulations show that word processing is facilitated by maximally contrastive inflectional systems, where verb forms exhibit the earliest possible point of lexical discrimination. Conversely, word learning is favored by a maximally generalizable system, where forms are inferred from the smallest possible number of their paradigm companions. Based on evidence from the literature and our own data, we conjecture that the resulting balance is the outcome of the interaction between form frequency and morphological regularity. Big families of stem-sharing, regularly inflected forms are the productive core of an inflectional system. Such a core is easier to learn but slower to discriminate. In contrast, less predictable verb forms, based on alternating and possibly suppletive stems, are easier to process but are learned by rote. Inflection systems thus strike a balance between these conflicting processing and communicative requirements, while staying within tight learnability bounds, in line with Ackermann and Malouf's Low Conditional Entropy Conjecture. Our quantitative investigation supports a discriminative view of morphological inflection as a collective, emergent system, whose global self-organization rests on a surprisingly small handful of language-independent principles of word coactivation and competition.

Keywords: morphological complexity, discriminative learning, recurrent neural networks, self-organization, emergence, processing uncertainty, stem-family size

Edited by:
Gonia Jarema,
Université de Montréal, Canada

Reviewed by:
Charalambos Themistocleous,
Johns Hopkins Medicine,
United States
Yu-Ying Chuang,
University of Tübingen, Germany

***Correspondence:**
Claudia Marzi
claudia.marzi@ilc.cnr.it
Marcello Ferro
marcello.ferro@ilc.cnr.it

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *09 December 2018*
Accepted: *19 August 2019*
Published: *10 September 2019*

Citation:
Marzi C, Ferro M and Pirrelli V (2019) A Processing-Oriented Investigation of Inflectional Complexity. Front. Commun. 4:48. doi: 10.3389/fcomm.2019.00048

1. INTRODUCTION

Assessment of the complexity of the inflection system of a language and its comparison with a functionally-equivalent system of another language are hot topics in contemporary linguistic inquiry (Bearman et al., 2015). Their goals may vary from a typological interest in classifying different morphologies, to a search for the most compact formal description of an inflection system, to an investigation of the nature of word knowledge and its connection with processing and learning issues (Juola, 1998; Goldsmith, 2001; Moscoso del Prado Martín et al., 2004; Bane, 2008; Ackerman and Malouf, 2013).

From a cross-linguistic perspective, the way morphosyntactic features are contextually realized through processes of word inflection probably represents the widest dimension of grammatical variation across languages, in a somewhat striking contrast with universal invariances along other dimensions (Evans and Levinson, 2009). This has encouraged linguists to focus on differences in morphological marking. Inflectional complexity is thus approached as a problem of feature counting, through comprehensive catalogs of the morphological markers and patterns in a given language or language type. In contrast with such "enumerative" approaches (Ackerman and Malouf, 2013), information-theoretic models have addressed the issue in terms of either algorithmic complexity (Kolmogorov, 1968), measuring the length of the most compact formal description of an inflection system, or in terms of information entropy (Shannon, 1948), which measures the amount of uncertainty in inferring a particular inflected form from another form, or, alternatively, from a set of paradigmatically related forms.

An altogether different approach, more conducive to addressing fundamental psycholinguistic and cognitive issues, is to conceive of complexity as related to the problem of *learning how to process* an inflection system. This step has far reaching consequences for the way we look at word knowledge, shifting our focus from what speakers know when they know inflection (mainly representations), to how speakers develop knowledge of inflection by processing input data (learning). According to this perspective, redundant patterns are predominantly statistical, and even irregularities appear to be motivated by their frequency distribution in the system and by the speaker's processing bias. All these issues are very different in character from the formal constraints on units, representations or rule systems proposed within theoretical and computational models, and make room for an empirical validation grounded in learning theory.

In this paper, we show the potential of such a learning-based, processing-oriented view of inflection complexity through a quantitative analysis of the behavior of Temporal Self-Organizing Maps (or *TSOM*s, Ferro et al., 2011; Marzi et al., 2012; Pirrelli et al., 2015), a recurrent variant of classical Self-Organizing Maps (Kohonen, 2002), independently trained on the inflectional systems of Standard Modern Arabic, English, German, Greek, Italian and Spanish.

The choice of *TSOM*s as neuro-biologically inspired computational learning models is motivated here by practical and theoretical reasons. First, their role is instrumental, as we use them to illustrate the dynamic approach to complexity we intend to advocate. Hence, some of the points made here will also hold, with some qualifications, for other existing computational models (Althaus and Mareschal, 2013; Baayen et al., 2019; Li et al., 2007; Mayor and Plunkett, 2010, among others). Secondly, *TSOM*s are based on discriminative principles of selective synchronization of processing nodes, supporting a *Word-and-Paradigm* view of the mental lexicon[1]. Accordingly, words are dynamically represented as emergent, superpositional patterns of short-term node activation, reflecting gradient levels of lexical specificity: from holistic to decompositional lexical representations. Thirdly, for each input word, *TSOM*s make it possible to inspect levels of node activation during online processing with a fixed sampling rate. This allows us to monitor patterns of node activation changing non-linearly with time as more symbols of the input word are presented, and check how levels of uncertainty in processing correlate with structural transitions in the input word (e.g., from a stem to its inflectional ending). Finally, we can correlate the *TSOM* temporal patterns with multiple defining features of typologically different inflectional systems, including: (i) the number of realization patterns of formal contrast (ranging from suppletion to extensive syncretism, through a whole range of intermediate cases); (ii) the type of such formal patterns (continuous, discontinuous or mixed); (iii) the amount of formal transparency they exhibit (ranging from a more fusional to a more agglutinative pattern); (iv) the frequency distribution of inflectionally marked forms within the same paradigm, and their relative dominance; (v) the frequency distribution of markers in their inflectional classes, and (vi) the amount of interpredictability among patterns (e.g., how easy it is for a speaker to predict an unknown inflected form from an already known inflected form of the same paradigm) as a function of their transparency and systematic nature.

As we shall see, data from our simulations makes room for an insightful analysis of time-bound patterns of structural organization of the network and its processing responses, accounting for the correlation between the complexity of inflectional data and processing aspects like speed of recognition, learnability and ease of production. Variation in complexity may have multiple, and in some cases opposite effects on all these issues, depending on the factors involved. For example, while regularly inflected forms are easier to learn because they take part of larger paradigm families, irregularly inflected forms are easier to discriminate because they are more isolated and ultimately less confusable. Based on non-linear analyses of simulation data

[1] *Word-and-Paradigm* Morphology (Matthews, 1991; Blevins, 2016) explicitly avoids the central problem of segmenting an inflected form like *walked* into meaningful sublexical parts (or "morphemes," e.g., *walk-ed*), and expressing the morpho-lexical content of the whole as a function of the morpho-lexical content of its parts. Rather, inflected words are themselves parts of two parallel networks of formal and semantic relations between members of the same paradigm. Crucially, Word and Paradigm Morphology says nothing about the ways the two networks are mutually related, thus avoiding problems of morphemes with no meanings (empty morphemes), meanings with no morphemes (zero morphemes), bracketing paradoxes etc. This strikes us as the most typologically-neutral approach to an assessment of inflectional complexity, as it eschews any attempt to put a morphemic straitjacket to non agglutinative languages. As *TSOM*s are trained on inflected forms with no morphosemantic and morphosyntactic content, only redundant formal relations are learned.

reported in this paper and other linear analyses from our previous work (Marzi et al., 2018), we suggest that inflection is better understood as a dynamic, potentially unstable system, whose complexity results from a balancing act between competing processing and communicative requirements, finely tuned during language acquisition.

2. INFLECTIONAL COMPLEXITY

Inflection lies at the intersection of two independent but interacting issues: (i) what syntactic contexts require morphosyntactic and/or morphosemantic word marking, and for what lexical and grammatical units; (ii) how morphosyntactic and morphosemantic information is overtly realized on lexical and grammatical units. In this paper, our attention will be devoted to the second issue, by simulating and evaluating the learning dynamics of differently graded morphological (ir-)regularities. To investigate the issue of how degrees of inflectional complexity affect word processing strategies, we selected 6 languages that are differently positioned in a typological continuum ranging from a more isolating type to a more fusional inflecting one:[2] namely English, German, Spanish, Italian, Greek, and, to broaden our typological coverage, Standard Modern Arabic.

We will focus on systemic aspects of lexical organization that are common to most (if not all) inflection systems, and in particular we will examine the possibility of discovering morphological organization in the implicative network of relations between words, in order to understand the impact of this organization on word processing and learning issues. All inflection systems share the property that they have families of related word forms exhibiting collective properties that cannot be deduced from any one of these forms individually. This is the hallmark of inflectional paradigms, i.e., clusters of fully inflected forms which are associated with an individual lexical exponent (e.g., English *walk*, *walks*, *walking* and *walked*) and mutually related through inflectional classes (i.e., classes of inflectional markers in complementary distribution).

2.1. Paradigm Complexity

Over the last decades, investigation of the formal properties of inflectional paradigms has played a prominent role in changing the research agenda in inflectional morphology, marking a tendency to move away from part-whole relations within complex words, toward descriptive relations between inflected forms (Matthews, 1991; Blevins, 2016). Accordingly, the complexity of a paradigm resides in two basic dimensions. The first dimension defines the amount of full formal contrast realized within the paradigm, i.e., how many different markers are uniquely associated with distinct combinations of morphosyntactic features (or paradigm cells, e.g., present indicative 1st singular), and how structurally different these markers are. The second dimension describes the number of interpredictability patterns between words in the same paradigm, i.e., how easy it is for a speaker to infer an unknown paradigmatic form from other familiar forms within the same paradigm.

Both dimensions of complexity are functionally relevant. Patterns of formal contrast serve to distinguish paradigmatically-related forms, and associate them with specific cells to communicate meaning or function within an inflectional system. We refer to this dimension as the level of "discriminative complexity" of the system. At the same time, variation in patterns of formal contrast is not scattered randomly across paradigm cells. Formal patterns tend to be interdependent, to the extent that knowing the inflection in one paradigm cell allows speakers to predict the inflection in another cell. This type of interdependency defines the "inferential complexity" of an inflectional system: paradigms where many forms are predicted upon exposure to one or few forms only, are *less* complex than paradigms where fewer forms are predicted from other forms.

From a cross-linguistic perspective, the discriminative complexity of an inflectional system is typically measured by enumerating the number of category values instantiated in the system (e.g., person, number or tense features) and the range of available markers for their realization. Accordingly, the larger the number of paradigm cells and their markers, the more difficult the resulting system (McWorther, 2001; Bickel and Nichols, 2005; Shosted, 2006). This notion of "Enumerative Complexity," however, has been criticized on several grounds (Bane, 2008; Sagot and Walther, 2011; Ackerman and Malouf, 2013; Sagot, 2018). It soon gets difficult, if not impossible, to compare inflectional systems whose feature-value inventories differ dramatically. For instance, according to the World Atlas of Language Structures (Haspelmath et al., 2005), there are more than 10 different cases in Hungarian and none in Swahili, and more than 5 gender values in Swahili as opposed to none in Hungarian. In cases like this, we have no principled basis for arguing that either system is more or less complex than the other. In addition, even in the most favorable case of two hypothetical systems whose feature-value inventories match perfectly, simple counting can be misleading. Suppose we are comparing two systems, each with only two categories (say, singular and plural) and three different endings for each category: A, B, C for singular, and D, E, and F for plural. In the first system, paradigms are found to present three possible pairs of endings only: <A, D>, <B, E>, <C, F>, which can be described as corresponding to three different inflection classes. In the second system, any combination is attested. Clearly, the latter system would be more difficult to learn than the former, as it makes it harder to infer the plural form of a word from its singular form (or the singular from the plural for that matter). Nonetheless, both systems present the same degree of Enumerative complexity.

Recently, information theoretic approaches have been proposed as a way to circumvent the limitations of pure feature counting methods. Following Kolmogorov (Kolmogorov, 1968), the complexity of a dataset of inflected forms can be measured

[2] According to Dressler and colleagues (Bittner et al., 2003), European languages can be arranged along an inflectional complexity continuum, ranging from a more inflecting/fusional type (left) to a more isolating type (right):

Lithuanian→Greek→Russian→Croatian→Italian→
Spanish→German→Dutch→French→English.

by the shortest possible grammar needed to describe them, or their *algorithmic complexity*, in line with the Minimum Description Length (*MDL*) principle (Rissanen, 2007). For example, following Goldsmith (2001), we can model the task of morphological induction as a data compression problem: find the collection of markers forming the shortest grammar that fits the empirical evidence best. To illustrate, English conjugation can be modeled as consisting of sets of endings in complementary distribution (somewhat reminiscent of "inflectional classes," referred to by Goldsmith as "signatures"), and a set of stems. For example, the list <*NULL*, *-er*, *-ing*, *-s*> is a signature for the verb stems *count*, *walk* and *mail*, but not for *love* or *notice*, which require the stems *lov-* and *notic-*, and the signature <*e*, *-ed*, *-ing*, *-es*>. Licensing an irregular verb like *drink* in this grammar formalism is even more verbose, as it requires three stems, *drink*, *drank* and *drunk*, and two signatures: <*NULL*, *-ing*, *-s*> for *drink*, and <*NULL*> for both *drank* and *drunk*. Goldsmith's algorithm, however, models paradigm learning as a top-down optimization problem, boiling down to a grammar evaluation procedure. The segmentation of forms into sublexical constituents is based on heuristic criteria and makes no contact with the problem of finding the minimally redundant set of paradigms. Ultimately, we are left with no clues about how word processing (segmentation) interacts with the paradigmatic organization of the morphological lexicon.

An alternative information-theoretic approach to complexity is based on Shannon's entropy (Shannon, 1948), or *informational complexity* of a set of paradigms. Information complexity rests on the intuition that a more complex system of inflected forms presents fewer interpredictability patterns between existing forms than a less complex system does. Ackerman et al. (2009) and Ackerman and Malouf (2013) use Shannon's information entropy to quantify prediction of an inflected form as a paradigm-based change in the speaker's uncertainty. They conjecture that inflectional systems tend to minimize the average conditional entropy of predicting each form in a paradigm on the basis of any other form of the same paradigm (the Low Conditional Entropy Conjecture, or LCEC). This is measured by looking at the distribution of inflectional markers across inflection classes in the morphological system of a language. More recently, Bonami and Beniamine (2016) propose to generalize affix-to-affix inference to inference of intraparadigmatic form-to-form alternation patterns (Pirrelli and Yvon, 1999; Albright, 2002). This approach offers several advantages. It avoids the need for the theoretically-loaded segmentation of inflected forms into stems and affixes in the first place. Secondly, it models implicative relations between stem allomorphs (or stem-stem predictability), thereby providing a principled way to discover so-called "principal parts," i.e., a minimal set of selected forms in a paradigm from which all other paradigm members can be deduced with certainty (Finkel and Stump, 2007). Finally, it emphasizes the role of joint prediction, i.e., the use of sets of forms to predict one missing form of the same paradigm, as a convenient strategy to reduce the speaker's uncertainty in addressing the cell filling problem.

In spite of their differences, however, both Kolmogorov's and Shannon's approaches are biased by a few *a priori* assumptions. Results obtained from the use of algorithmic complexity strongly depend on the formalism adopted for grammatical description (Sagot, 2018): for example, surface representations of verbs are typically segmented into stem-ending patterns. Information entropy dispenses with segmentation, but it rests crucially on the types of formal relations used to identify predictability patterns in inflectionally related forms. In either case, the algorithm does not adapt itself to the specific structural requirements of the systems it analyzes. Instead, it presupposes considerable knowledge of the morphology of the target language, in order to assess how effectively that knowledge can describe the language.

A more principled approach to measuring morphological complexity is to investigate the impact of incremental learning and online processing principles on paradigm organization, based on observation of the behavior of an unsupervised algorithm acquiring an inflection system from fully-inflected forms. The approach is central to establishing a connection between human language behavior, word distributions in input data, learning mechanisms and the taxonomy of units and combinatorial principles of linguistic theories. Careful quantitative analysis of the way a computational learning system can bootstrap structural information from typologically different training sets is not too far from what is done in experimental psycholinguistics, where the role of multiple factors on human processing is investigated by controlling factor interaction in the execution of a specific processing task. In the end, this allows us to frame the problem of inflectional (and paradigm) complexity into the larger context of word processing complexity, to which we turn in the following section.

2.2. Processing-Oriented Complexity

The discriminative and inferential dimensions of paradigm complexity illustrated in section 2.1 meet potentially competing communicative requirements. A maximally contrastive inflection system is one where inflected forms, both within and across paradigms, present the *earliest possible point of recognition*[3], i.e., the position where they are uniquely distinguished from their paradigm companions. From this perspective, extensive suppletion (e.g., with English *be/am/are/is/was/were*) reflects a recognition-driven tendency for a maximally efficient contrast.

A maximally contrastive system, however, may require extensive storage of its forms, insofar as relatively few items can be *inferred* from its paradigm companions. Hence, any such system is not only slow to learn, but also fairly demanding and inefficient to use. Due to the Zipfian distribution of the forms in use within a speech community, almost half of the word forms of a language occur only once in a corpus, irrespective of corpus size (Blevins et al., 2017). This means that even high-frequency paradigms will tend to be partially attested, and speakers must be able to generalize available knowledge to rare events if they want to interpret or produce novel forms.

Issues of morphological regularity also have a bearing on the relationship between morphological complexity and word processing. We already noted that intra-paradigmatic suppletion,

[3]Starting from Marlsen-Wilson's classical definition of "uniqueness point" (Marslen-Wilson, 1984), we will make use of more recent refinements of this notion (Balling and Baayen, 2008, 2012).

arguably the most radical break with systematic and predictable inflection, is functional in maximizing contrast, but may represent a hurdle for word learning. Conversely, syncretic realization of many paradigm cells with identical forms (as in the English present indicative sub-paradigm) and, more generally, paradigm ambiguity, can slow down acquisition of overtly marked forms (e.g., the third singular present indicative *s*-forms in English). Researchers from diverse theoretical perspectives observe that rich inflection in fact facilitates early morphological production. In competition-based (Bates and MacWhinney, 1987), as well as functional (Slobin, 1982, 1985) and cue-response discriminative perspectives (Baayen et al., 2011), non-syncretic morphological paradigms such as those of Italian conjugation are argued to provide better syntactic cues to sentence interpretation, as compared, for example, to the impoverished inflectional system of English verb agreement. Biunique form-meaning relationships make inflectional markers more transparent, more compositional and in the end more easily acquired than the one-to-many mappings of morphological forms to syntactic features that are found in English, Swedish and Dutch (Phillips, 1996, 2010). Some researchers (Crago and Allen, 2001; Blom, 2007; Legate and Yang, 2007; Xanthos et al., 2011) have focused on the amount of finite verbs that children receive from the adult input, to observe that the high percentage of overtly inflected forms correlates with the early production of finite forms by children. In the framework of Natural Morphology, Dressler and colleagues (Bittner et al., 2003) claimed that a richer inflectional system makes children more aware of morphological structure, so that they begin to develop intra-paradigmatic relations sooner than children who are confronted by simpler systems do (Xanthos et al., 2011).

The literature reports a number of effects that family size and the frequency of family members have on a variety of processing tasks. A large family size supports visual word recognition, with printed words with many neighbors being recognized more quickly than words with fewer neighbors (Andrews, 1997). However, when neighbors are considerably more frequent than the target word, recognition of the target is inhibited. A similar reversal from facilitation to inhibition is reported in spoken word recognition and related tasks (Luce and Pisoni, 1998; Magnuson et al., 2007), where many neighbors are found to delay recognition of the target word. The effect has recently been interpreted as due to serial (as opposed to parallel) processing (Chen and Mirman, 2012): a word supported by a dense neighborhood is produced and read faster. But when the same word is presented serially (e.g., in spoken word recognition), high-frequency neighbors engage in competition and inhibit processing. Such a modality-driven subdivision of processing strategies is however not clear cut.

Balling and Baayen (2012) provide evidence for the combined processing effects of two uniqueness points in both auditory and visual processing. The first is the word's initial Uniqueness Point (*UP1*), where an input word is distinguished from its morphologically-unrelated competitors (e.g., *carrier* and *carpenter*). The second is the later Complex Uniqueness Point (*CUP*), where an input word is distinguished from its morphologically-related competitors (e.g., *writes* and *writing*). They report that late *UP1* and *CUP* are inhibitory in auditory and visual lexical decision, both independently and in interaction. In line with this evidence, more regular paradigms are predicted to favor entrenchment of shared stems and quicker acquisition of full forms, but they may cause larger effects of processing uncertainty at the stem-ending boundary than more irregular paradigms do, due to the larger range of possible following inflectional endings in regulars, as compared to irregulars.

The human exquisite sensitivity to word frequency distributions appears to be at the root of entropy-based processing effects. Milin et al. (2009a,b) report that speakers engaged in a lexical decision task are sensitive to the divergence between the word frequency distribution within a single paradigm, and the cumulative word frequency distribution in the inflection class to which the paradigm belongs. In particular, if an inflected form occurs in its paradigm less frequently than one would expect from the frequency of its ending, the visual recognition of that form gets more difficult. This finding has been confirmed in follow-up studies (Kuperman et al., 2010; Baayen et al., 2011; Milin et al., 2017).

Ferro et al. (2018) replicate these effects with *TSOM*s trained on inflectional paradigms with frequency distributions of varying entropy. The quantitative analysis of computer simulations is amenable to an interesting interpretation of entropic effects in terms of competition between family members. On average, word recognition is facilitated when words belonging to the same paradigm compete on a par, i.e., when their distribution is highly correlated with the distribution of their inflectional endings. When this is not the case, the entropy of the family decreases, increasing the risk that a low-frequency form is inhibited by a high-frequency member of the same family.

Marzi et al. (2018) add a cross-linguistic perspective on this evidence, by measuring the processing costs incurred by *TSOM*s trained on the inflectional systems of 6 languages. They observe that the averaged word processing costs for each language oscillate linearly within a small range (in keeping with Ackermann and Malouf's *LCEC*), whose upper and lower bounds are marked by Modern Greek conjugation (harder) and English conjugation (easier) respectively. All other inflectional systems present no statistically significant differences in the processing "overheads" they require, in spite of their typological diversity and their varying levels of morphological transparency. The analysis supports an interpretation of the *LCEC* as resulting from a balancing act between word processing and word learning. Word processing puts a premium on a system where word forms are maximally distinct and accessed early. Word learning, however, somewhat counteracts such a processing bias. Acquiring new forms is in fact necessary for speakers to communicate and discriminate in an ever changing social environment. The need to keep the system open thus increases processing uncertainty, as it inevitably increases the number of elements in the system.

Our present investigation is intended to go beyond these preliminary results. In the following sections, we investigate the competitive dynamic between word processing and word learning in more detail, focusing on the role that paradigm contrast plays in a variety of training conditions. Based on the data of Marzi et al. (2018), we model non-linear patterns of node activation in the online processing of inflected forms, in order to check how levels of processing uncertainty correlate with structural transitions in regularly and irregularly inflected forms within a single inflectional system, and across different inflectional systems. In addition, we model the impact of these effects on word learning, and we investigate the role that developing patterns of morphomic redundancy play during learning. As non-linear modeling can fit observed data patterns without *a priori* assumptions on the shape of the regression lines, this analysis is intended to lend further support to the dynamic view of inflectional complexity advocated here. In particular, we gain a better understanding of the basic language-invariant principles accounting for the discriminative and inferential dimensions of paradigm organization.

3. MATERIALS AND METHODS

In this section and in section 4, we provide data-driven evidence of the interaction between morphological complexity and word processing in six languages. The evidence is gathered through repeated computer simulations of the acquisition of the verb inflection system in each of the input languages. For this purpose, we use Temporal Self-Organizing Maps (*TSOM*s), a recurrent variant of Kohonen's Self-Organizing Maps (Kohonen, 2002) that offers a neurally-inspired computational model of *discriminative* learning of time-series of symbolic units (Ferro et al., 2011; Marzi et al., 2014; Pirrelli et al., 2015)[4]. *TSOM*s can learn all training sets accurately (**Table 1**), while showing, at the same time, differential effects in paradigm self-organization, learning pace, and word processing. We start this section with a short description of their architecture and learning principles. A more technical, detailed description is found in the **Appendix** to this paper.

3.1. Recurrent Topological Networks for Word Learning

In *TSOM*s, word learning proceeds by developing maximally discriminative Markov-like chains of topologically arranged memory nodes, derived from exposure to fully inflected input forms with no morphological annotation. Nonetheless, node chains can mirror effects of gradient morphological structure and emergent paradigm organization. By developing specialized patterns of map nodes through recurrent connections, *TSOM*s encode input symbols auto-associatively, exploiting the formal redundancy of symbolic temporal series. Node specialization is modeled through discriminative learning equations (Ramscar and Yarlett, 2007; Baayen et al., 2011), which offer a powerful strategy for scaffolding the input stream into internalized structured representations that are efficient for word recognition and production. We provide here an informal, functional description of *TSOM*s. The interested reader is again referred to the **Appendix** for all mathematical and algorithmic details.

[4] A running version of the *TSOM* package can be downloaded at http://www.comphyslab.it/redirect/?id=frontiers2019.

A lexical *TSOM* consists of a bank of input nodes encoding input letters, and a bank of processing nodes making up the lexical map proper (**Figure 1**). Each processing node is connected to the input layer through one-way input connections, and to other processing nodes (including itself) through one-time delay re-entrant temporal connections. At each time tick t, activation flows from the input layer to the map nodes through input connections. Re-entrant temporal connections update each map node with the state of activation of all nodes at the previous time tick (see the unfolded view of **Figure 1**). As with classical Recurrent Neural Networks (Elman, 1990), a word is input to a *TSOM* one symbol S at a time. Activation spreads through both input and temporal connections to yield an overall activation state, or Map Activation Pattern for S at time t: $MAP_t(S)$. The node with the highest activation level in $MAP_t(S)$ is called the Best Matching Unit for S at time t, or $BMU_t(S)$. A time series of sequentially activated *BMUs* will be referred to as a *BMU* chain below. It represents the map's cumulative memory trace for an input time series.

Weights on temporal connections encode how strongly the current BMU_t is predicted by BMU_{t-1}, over a continuous range from 0 to 1. Temporal connection weights are trained on input data according to the following principles of discriminative learning, strongly reminiscent of Rescorla and Wagner (1972) equations (see **Appendix**). When the bigram 'AX' is input, a *TSOM* goes through two learning steps:

1. the temporal connection between $BMU_{t-1}(A)$ and $BMU_t(X)$ is strengthened (*entrenchment*);
2. all other temporal connections to $BMU_t(X)$ are weakened (*competition*).

Step 1. and 2. incrementally enforce node specialization. Over the course of training, the map tends to allocate maximally distinct processing nodes, as a function of the form overlap and form frequency of input strings in the training set.

To illustrate the effects of node specialization, **Figure 2A** sketches two possible end-states in the allocation of *BMU* chains responding to 9 form types of German *geben* 'give': *geben* (infinitive, 1p and 3p, present indicative), *gibst* (2s, present indicative), *gibt* (3s, present indicative), *gebe* (1s, present indicative), *gab* (1s, 2s and 3s, preterite), *gaben* (1p and 3p, preterite), *gabt* (2p, preterite) and *gebend* (present participle). In the left panel, *BMUs* are arranged in a word tree. Any node N keeps track of all nodes that were activated in order to arrive at N. The right panel of **Figure 2A**, on the other hand, offers a compressed representation for the 9 form types, with shared letters activating identical *BMUs*. As a result, when the shared node 'b' is activated, one loses information encoding which node was activated at the previous time tick. Let us briefly consider the implications of the two structures for word processing. In the word-tree (left), *gibt*, *gabt*, and *gebt* are mapped onto distinct node chains, bifurcating at the earliest branching point in the hierarchy (after the "g" node). From that point onward, the three forms activate distinct node paths. Clearly, when a node has only one outgoing connection, the *TSOM* has no uncertainty about what step to take next, and can anticipate the upcoming input symbol with certainty. In the word-graph on the

TABLE 1 | Statistics for the 6 datasets.

Language	Min/max form length	Regular/irregular Paradigms	Form types/ Training size	Recoding %	Recall %	*sd* %
Arabic	4/11	18/28	560/601	100	99.93	0.16
English	2/11	20/30	208/750	100	99.62	0.86
German	3/11	16/34	504/750	100	99.76	0.18
Greek	2/13	37/13	744/750	100	99.84	0.06
Italian	2/12	23/27	748/750	100	99.79	0.15
Spanish	2/15	23/27	715/750	100	99.94	0.13

Form length is given by number of input symbols (with the exclusion of # and $), with orthographically marked stress being encoded as a distinct symbol. Differences between cardinality of form types and cardinality of the training set are due to syncretism. Percentage values of correctly recoded and recalled word types (and for the latter standard deviations) are given for each language, averaged over 5 map instances.

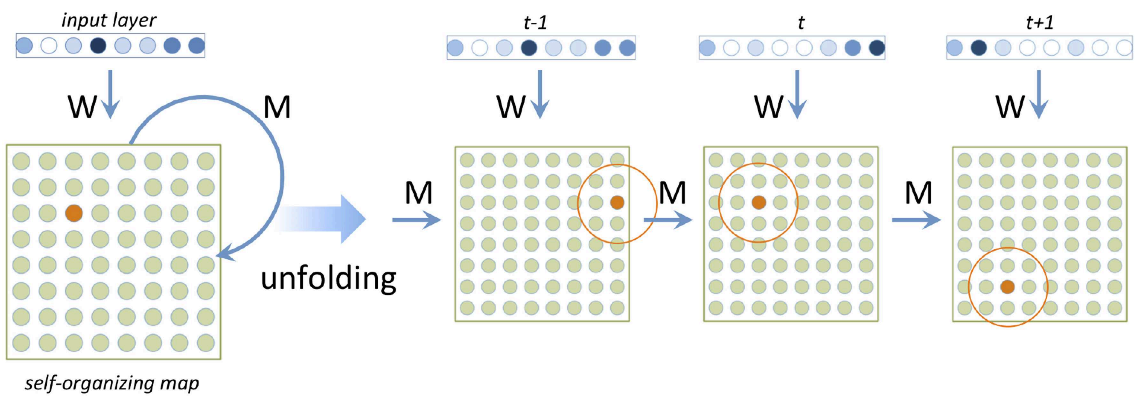


FIGURE 1 | Architecture and unfolded view of a *TSOM*. Orange nodes represent Best Matching Units activated at successive time ticks.

right, on the other hand, branching paths converge to the same node as soon as input forms share an identical symbol. Having more branches that converge to a common node increases processing uncertainty. The node encapsulates information of many preceding contexts, and its possible continuation paths are multiplied accordingly.

In **Figure 2B**, a word tree of the German verb *glauben* ("believe") is given for the 9 form types filling the same 14 paradigm cells selected for *geben* (**Figure 1A**). It is instructive to compare these structures. Of particular interest is the base stem *glaub*-, which is systematically shared by all its verb forms, and followed by a longer stretch of inflectional markers (conveying features of tense, person and number) than the base stems of *geben* (namely, *geb*-, *gib*- and *gab*-) are. Accordingly, the left-to-right processing of regularly inflected forms of *glauben* requires the traversal of more branching structures (and thus more indecision points). This marks an important structural difference between regular and irregular paradigms in all languages considered here, with a significant impact on the processing behavior of *TSOM*s, as we shall see later in more detail (section 5.1.1).

3.2. The Data

The *TSOM* architecture of **Figure 1** was used to simulate the acquisition of the verb inflection system of six languages: Standard Modern Arabic, English, German, Modern Greek, Italian and Spanish. For all languages, we selected 15 forms (14 forms for Arabic) for each of the 50 top frequency paradigms sampled from a reference corpus[5]. For each language, forms were chosen from a fixed set of paradigm cells. To the greatest extent possible, the set of cells was kept comparable across languages[6]. Each input form was transliterated into an ASCII-based sequence of possibly complex symbols preceded by the start-of-word symbol ("#"), and followed by the end-of-word symbol ("$"). For Modern Standard Arabic forms we used the Buckwalter transliteration rules[7].

[5]Our data sources are: CELEX (Baayen et al., 1995), for German and English; the Italian Paisà Corpus (Lyding et al., 2014); the European Spanish Subcorpus of the Spanish *TenTen* Corpus (www.sketchengine.co.uk); the SUBTLEX-GR corpus (Dimitropoulou et al., 2010) for Modern Greek, and the Penn Arabic Treebank (Maamouri et al., 2004). The full set of word forms, for each language, is available at http://www.comphyslab.it/redirect/?id=lrec2018_data.

[6]For English, German, Modern Greek, Italian and Spanish, the list includes all present and past indicative cells. English, German, Italian and Spanish also include the infinitive, past participle and gerund/present participle cells, which are replaced by the three singular cells of the simple future in Modern Greek. For Standard Modern Arabic, we selected the first, second and third masculine singular and plural cells, and the third feminine singular cell for both the imperfective and perfective.

[7]www.qamus.org/transliteration.htm.

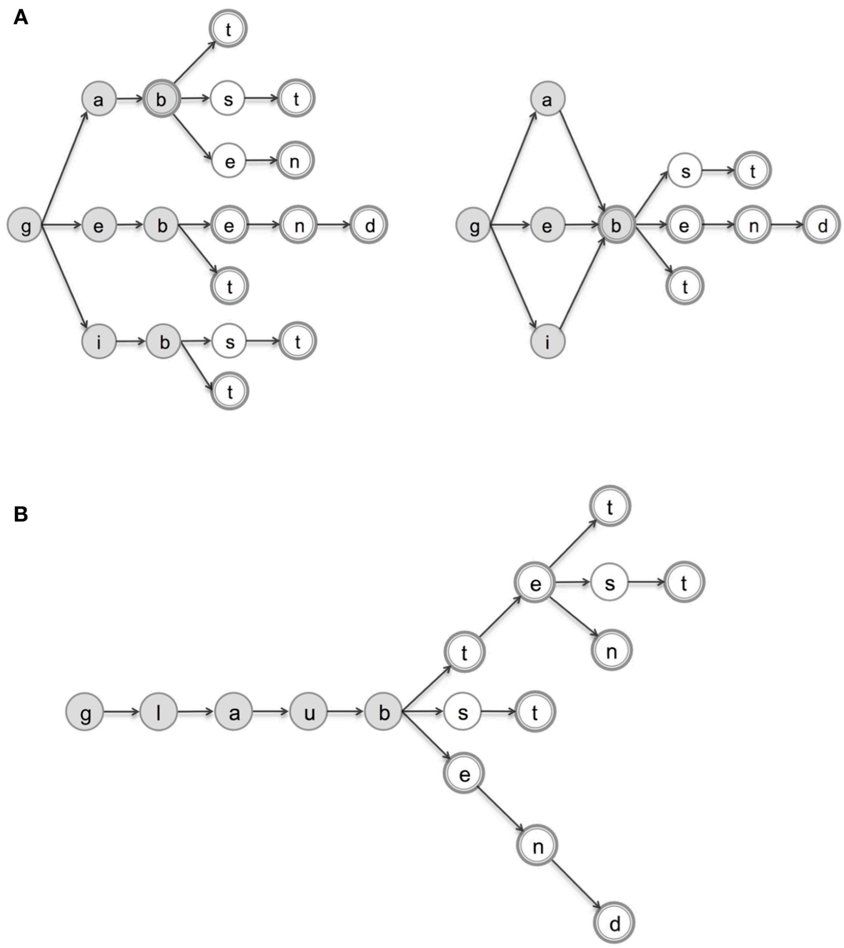


FIGURE 2 | Graph-based representations for a few German verb forms of *geben* "give" **(A)** and *glauben* "believe" **(B)**. Double-circled nodes represent word-final states, i.e., letters ending a word form; shaded nodes represent (possibly allomorphic) stems, according to the segmentation criteria of section 3.2. In **(A)**, we provide a word-tree (left) and a word-graph (right) for the same set of 9 form types of *geben*.

Although forms are input to the map with no information about their internal structure, the dataset was annotated manually, to allow us to correlate variation in the map response after training with variation in the (morphological) structure of training forms. First, for each form, we marked up two uniqueness points: *UP1* and *CUP*. *UP1* was marked at the position in the verb form where the stem is uniquely identified, both relative to *paradigmatically-unrelated* onset-aligned forms in the dataset, and with respect to other possible stem allomorphs *within the same paradigm*. For example, the stem *drunk* is distinguished from the unrelated *doubt* at "r" in second position, but it is distinguished from *drink* or *drank* only when "u" is reached in third position[8]. The marking of *CUP* followed Balling and Baayen in marking the point at which the whole verb form is distinguished from other forms of the same paradigm.

In addition, we based morphological segmentation on a uniform (PREFIX) + STEM + (SUFFIX) template, where the notion of verb stem has the Aronovian, morphomic status of an unpredictable allomorph of the base stem, conveying systematic, but possibly non homogeneous clusters of morpho-lexical features (Aronoff, 1994). Accordingly, formally predictable stems derived from their base stem through a systematic process of stem formation (e.g., English past(*walk*) → *walked*) are segmented into a base and a suffix (*walk-ed*). Inflected forms containing unpredictable stems (e.g., English past_part(*drink*) → *drunk*, German past_1S(*geben*) → *gab*, German pres_2S(*geben*) → *gibst*) are segmented into a stem and a suffix (whenever overtly realized) (*e.g., drunk-*, *gab-*, *gib-st*).

[8]In fact, Ballling and Baayen's definition of UP1 is intended to index competition from morphologically unrelated words only (Balling and Baayen, 2008, 2012). However, this is justified by the fact that their database only includes regularly inflected forms, i.e., inflected forms showing no stem allomorphy: in this case, identification of the stem automatically entails identification of the entire paradigm. Our operational definition of UP1 represents a natural extension of Balling and Baayen's criteria to a dataset containing irregularly inflected forms.

Paradigms that undergo unpredictable stem formation processes are classified as *irregular*. Paradigms whose stems are formed transparently and systematically[9] are annotated as *regular*. In Standard Modern Arabic, where verbs appear to cluster into different inflectional classes depending on the vocalic patterns intercalated with the roots, irregular (or weak) verb paradigms have stems that undergo a reduction in the number of root consonants (traditionally referred to as first, second, third weak and doubled roots, depending on the position and number of the radical consonants being dropped) (Marzi et al., 2017). Our segmentation template thus cuts across the traditional regular vs. irregular classification, yielding *ya-ktub-u* "he writes" and *katab-a* "he wrote" (as an example of a regular paradigm), together with *ya-qūl-u* "he says" and *qāl-a* "he said" (as an example of an irregular paradigm). Finally, for those languages (like Modern Greek and Italian) where regular paradigms exhibit unpredictable stem vowel selection (Ralli, 2005, 2006), stem identity/regularity was based on the form obtained by dropping the thematic vowel from the stem (Pirrelli, 2000; Pirrelli and Battista, 2000; Bompolas et al., 2017).

To annotate graded regularity in our data, we define, for each paradigm, the stem-family of a target form as the set of its formally distinct stem-sharing members. For instance, in our English training set, the forms *drink*, *drinks* and *drinking* are each associated with a stem-family of size 2, whereas *drank* and *drunk* have each an empty family. Likewise, the Italian verb forms *veng-o* "I come" and *veng-ono* "they come" have each a stem-family of size 1, as do *vien-i* "you come" (2s) and *vien-e* "(s)he comes." In contrast, *ven-ire* "to come," *ven-iamo* "we come," *ven-ite* "you come" (2p), *ven-isti* "you came" (2s), *ven-immo* "we came," and *ven-iste* "you came" (2p) have a stem-family of size 5. Finally, for each paradigm, we calculate its average stem-family size, and then normalize it by dividing it by the maximum number of possible stem-sharing members in the paradigm. Accordingly, the paradigm regularity for *drink* is $\frac{2+2+2+0+0}{5} \cdot \frac{1}{4} = \frac{3}{10}$. The score, which we will refer to in section 4 as *gradient paradigm regularity*, ranges between 0 (minimum value obtained when each paradigm member is formed on a distinct stem) and 1 (maximum value, obtained when a paradigm is fully regular, i.e., when the stem-family of each paradigm member includes all other members of the same paradigm).

We used the same setting of free learning parameters for training *TSOM*s on all languages. However, to minimize the impact on the map topology of cross-linguistic differences in the training data (for instance, in terms of form length and number of form types: see **Table 1**)[10], the number of nodes for each language-specific map was made to vary so as to keep constant the ratio between the map size and the number of nodes required to represent its training set with a word tree. Accordingly, a language-specific map size ranges between 35 × 35 nodes (English) and 42 × 42 nodes (Greek). Due to the combined effect of keeping this ratio constant across languages and preventing the temporal neighborhood radius (see **Appendix**) from decreasing to zero, maps tend to develop overlapping node chains for word forms sharing the same ending. Ultimately, these constraints on the topological organization of the map avoid data overfitting, as the map cannot possibly build up a dedicated memory traces for each form in the training set, i.e., it cannot memorize the original input data in its entirety (Marzi et al., 2014). For all languages, we stopped training at epoch 100, when all learning parameters reach minimum plasticity (see **Appendix**). To control for random variability in the map response, for each language we trained and tested the map 5 times on the same data, and averaged the results across all iterations.

4. RESULTS

To analyze the results of our simulations, we focused on the way *TSOM*s *process* input words, and adjust their processing strategies while *learning* different inflectional systems.

Word processing describes the map's short-term response to an input word. The response consists in a distributed pattern of node activation and includes three sub-processes: *input recoding*, when a specific winning node (a *BMU*) is activated by an indvidual stimulus (a letter); *prediction*, i.e., the map's on-line expectation for an upcoming input letter, after a time series of input letters is presented on the input layer; and *recall*, which consists in retrieving a series of letters from the map's response to the series (the corresponding activation pattern). In turn, *word learning* consists in associating systematic, long-term patterns of node activation with input sequences of letters, so that each sequence can be discriminated from other sequences (i.e., it elicits a distinct response from the map), and letters can be retrieved from the pattern in their appropriate order for word *recall*. Accordingly, we can define the time of acquisition of a word as

[9]In stem formation, transparency and systematicity do not necessarily correlate with predictability. Modern Greek provides a good example of transparent and systematic stem formation processes which however are not predictable, as they require a thematic vowel whose selection is based on lexical information (Ralli, 2014; Bompolas et al., 2017).

[10]**Table 1** shows that English and (to a lesser extent) German conjugations exhibit identical verb forms (morphological syncretism) in a sizable subset of the 15 paradigm cells we selected. It can be objected that we are imposing the paradigm structure of inflectionally richer languages on inflectionally more impoverished ones, with the effect of spuriously inflating the token distribution of our dataset in languages such as English, where the base form is presented to the TSOM five times as many as $-s$ forms are. However, we contend that our training sets are indeed comparable cross-linguistically, and can mirror realistic developmental conditions. The paradigmatic space of a language is not defined by the amount of morphological contrast conveyed by its inflected forms, but by the set of morphosyntactic relations (e.g., subject-verb agreement) holding between word forms in context. Children acquire inflected forms by binding them to larger contexts (e.g., *she's walking*, *it rains*, etc., Wilson, 2003; Pine et al., 2008), and by abstracting away from recurrent combinations of number, person, gender, case, tense and aspect features in context. If a specific inflected form is seen in different such combinations (e.g., *I walk*, *you walk*, *we walk*, *they walk*), the child is repeatedly exposed to an invariant form, whose token frequency will correspondingly increase relative to other paradigmatically-related forms (e.g., *walks* and *walking*). This is confirmed by developmental evidence showing that extensive syncretism has perceivable effects on the child's intake of the amount and distribution of morphological contrast exhibited by a language inflection (Legate and Yang, 2007; Pine et al., 2008; Xanthos et al., 2011; Krajewski et al., 2012). Thus, although we are deliberately partialing out token frequency effects (which appear to show complex a interaction with word category, lexico-semantic class, modality and age of acquisition: see Bornstein et al., 2004; Goodman et al., 2008), our data enable us to focus on non trivial type-driven aspects of dynamic paradigm self-organization.

the learning epoch when the map is able to consistently recall the word correctly from its activation pattern, while making no mistakes at later epochs. It should be emphasized that, from this perspective, learning simply consists in developing long-term processing patters from repeated, short-term successful processing responses. We will return to this mutual implication between short-term processing and long-term learning in the concluding section of this paper.

Accuracy of *recoding* measures the ability of a *TSOM* to correctly map an input word form onto a chain of *BMU*s. When presenting a *TSOM* with a time-series of symbols making up an input word, the word form is recoded correctly if all *BMU*s are associated with the correct input symbols in the appropriate order (see Equation 1 in **Appendix**). Accuracy of *recall* measures the ability of a *TSOM* to correctly retrieve a word form from its memory trace (or Integrated Activation Pattern, see **Appendix**). This is done by iteratively spreading node activation from the start-of-word node ("#") through the nodes making up the temporal chain of an input word. At each time step, the *TSOM* outputs the symbol associated with the currently most highly-activated node. The step is repeated until the node associated with the end-of-word symbol ("$") is output (see Equations 19–21 in **Appendix**).

By measuring the map's *prediction* rate, we assess the ability of a trained map to predict an incrementally presented input word. Prediction scores across input words are calculated by symbol position. The prediction score of a correctly predicted symbol is obtained by increasing the prediction score of its immediately preceding symbol by 1-point. Wrongly predicted symbols are given a 0-point score (see Equation 18 in **Appendix**). For each input word, the more symbols are predicted, the higher the prediction score assigned to the word. High prediction scores thus reflect strong expectations over upcoming input symbols, and measure the relation between successful serial word processing and accurate positional encoding of symbols in a time series.

Accuracy scores for recoding and recall are given in **Table 1**. The performance of each language-specific map is evaluated by averaging scores across five iterations of the learning experiment for each inflection system. Performance is consistently good for all the languages in our experiments, especially when we focus on the accuracy of recall, considering the difficulty of the task. This shows that the algorithm is adaptive enough to be able to fine tune its time-sensitive representations to the orthotactic and morphotactic redundancies of each system, irrespective of the specific position a system takes along the isolating-inflecting continuum.

Some interesting differences are observed when we consider the learning pace of the maps for the different languages. **Figure 3A** shows the boxplot distribution of per-word learning epochs for the 6 languages. In particular, Arabic per-word learning epochs are significantly later than those of all other languages ($p < 0.001$), with English per-word learning epochs being significantly earlier ($p < 0.01$). In **Figure 3B** we plotted, for each language, its paradigm *learning span*. The span measures the number of epochs taken by a map to learn an entire paradigm after the first member of the paradigm is learned. The span thus quantifies the average inferential ability of a map learning an inflection system, and, indirectly, the learnability of the system. The ranking order of the learning span for the languages of our experiments nicely mirrors the relative complexity of the inflectional processes in the six languages, according to Bittner et al. (2003). On the one hand, the learning span for the English verb set is significantly smaller than the learning span of all other languages ($p < 0.01$ compared with German, and $p < 0.001$ compared with all other languages). On the other hand, the learning span for the Greek verb set is comparatively larger than the span of the remaining languages ($p < 0.001$), followed by the Italian set (which is not significantly different from the Greek one), and the Spanish one. In fact, Greek, Italian and Spanish present, unlike English and German, a verb system with more conjugation classes. The inferential process of filling empty cells from one attested form of the same paradigm thus requires preliminary identification of the appropriate conjugation class, with its set of inflectional endings[11]. Nonetheless, the comparatively small range of variation (between 3 and 8 for median values) that these languages exhibit in their learning span lends support to Ackerman and Malouf's conjecture (Ackerman and Malouf, 2013) that the inferential complexity of inflectional system must oscillate within tight entropic bounds.

Based on the results reported in this section, it appears that *TSOM*s are capable of learning the underlying syntagmatic structure of inflection (consisting of serialized sublexical units like prefixes, stems and suffixes), and the inferential patterns necessary for generalizing this structure. Nonetheless, this is only partially satisfactory for our present concerns. We would like to know not only that a particular inflection system can be learned, but also how it can be learned, what is actually being learned, and how hard it is for a *TSOM* to learn the inflection system of a language compared with that of another language. To reach this level of understanding, in the following sections we report statistical analyses of the performance of *TSOM*s in the prediction task and investigate their learning dynamic by using the R software (R Core Team, 2014). Our statistical models will provide a careful, quantitative interpretation of what the map does, and will enable us to suggest a non impressionistic categorization of the time-bound representations hidden in the map's topological self-organization, and match up these categories with familiar linguistic notions.

5. DATA ANALYSIS AND INTERPRETATION

The internal structure of a *TSOM* is not easy to examine, but there are ways to correlate the structure and distribution of the training data with the functional behavior of the map (e.g., its ability to predict what is coming next at a certain point in time in the input based on what was input until that point). In this section, we first consider categorical features of the training data: namely the language being input and its inflectional syntagmatics, by

[11]This is not needed in Standard Modern Arabic, whose learning span is smaller than in Greek, Italian and Spanish (but larger than in English and German), as Arabic class information is, in most cases, already conveyed by the verb stem.

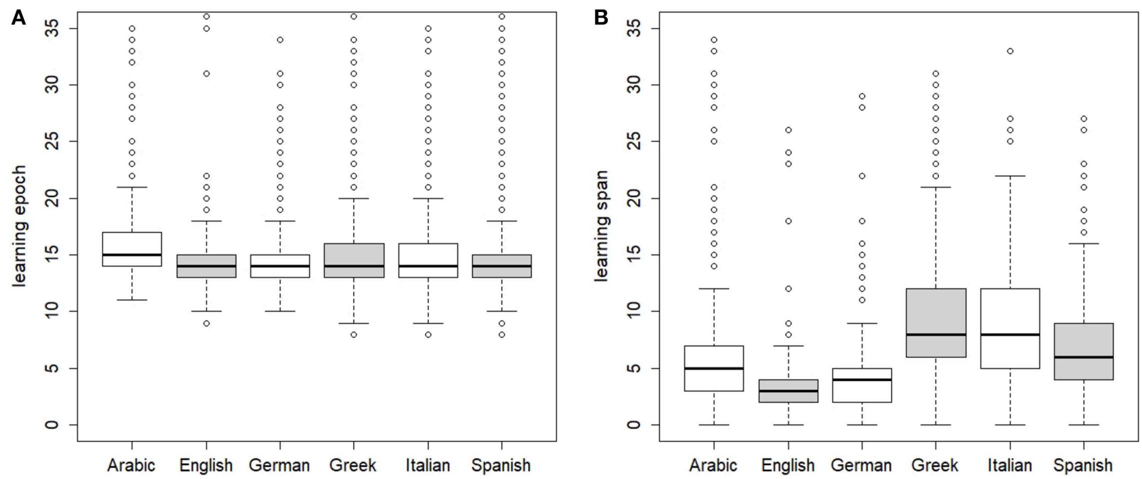


FIGURE 3 | For each language dataset, boxplot distributions of **(A)** per-word learning epoch and **(B)** per-paradigm learning span.

looking at the way sublexical constituents are annotated, and inflected forms are classified as regular or irregular (section 5.1). The idea is to explore the language-specific sensitivity that the map develops while being trained on a given language and its set of inflectional processes. Secondly, we move on to take a *Word-and-Paradigm* perspective on the same material, with a view to matching up the behavior of the map with a more graded classification of morphological regularity, and with more fundamental, cognitive effects of family size on the map's behavior (section 5.2.1). From both perspectives, we examine processing effects as well as learning effects.

5.1. Cross-Linguistic and Regularity Effects

5.1.1. Processing Effects

The leftmost panel of **Figure 4** plots the linear rate of letter-by-letter *prediction* when a trained *TSOM* is processing regularly (cyan line) or irregularly inflected forms (red line), as a function of letter position (distance) to the Morpheme Boundary (*MB*) between stem and suffix, where *MB* = 0 corresponds to the first element of the suffix. The prediction rate counts how many consecutive letters are predicted by the map during word processing. The plot is a linear regression model of the letter prediction rate, using as predictors the letter position and the dichotomous regular vs. irregular classification (see section 3.2). Verb forms from the 6 languages are considered cumulatively in the model.

Positive slopes indicate, unsurprisingly, that the map gets more and more accurate in anticipating an upcoming symbol as more of the word input is consumed. As has already been shown (see Marzi et al., 2018 for preliminary results on the same set of the 6 languages), *TSOM*s appear to be significantly more accurate in predicting forms belonging to regular paradigms than to irregular ones, but no apparent interaction is observed between regularity and distance to the stem-suffix boundary. To a first approximation, the evidence seems to show that inflectional regularity consistently facilitates word processing, independently of how inflection is marked cross-linguistically. In fact, the center panel of **Figure 4** shows that the 6 languages of our sample present slightly different prediction slopes. Arabic and English exhibit a steeper prediction slope compared with the slope of German, Greek, Italian, and Spanish.

A more interesting picture[12] emerges when we move from a linear regression model to a non-linear regression model of the same data. The rightmost panel of **Figure 4** shows how prediction rates vary across languages when the time course of word prediction is modeled by a Generalized Additive Model (*GAM*). *GAMs*, and related plots (obtained with the *ggplot* function), eliminate spurious linear leveling, and allow changes over serial processing to be modeled in a fine-grained way[13]. The ascending path for increasing values of *distance to MB* basically confirms the linear trend observed in the center panel of **Figure 4**. But the non-linear interpolation highlights an important discontinuity at the stem-suffix boundary (for *distance to MB*=0). In English, Greek, Italian and Spanish, the stem-suffix boundary marks a sharp drop in prediction rate, supporting the intuition that when the final letter of a stem is recognized, the map has to revise its expectations for an upcoming symbol. We interpret this as an effect of structural discontinuity between the stem and the following suffix, which is deeper in regularly inflected forms than it is in irregularly inflected forms for all languages in our sample (**Figure 5A**).

The same mechanism accounts for the fact that Greek, Italian and Spanish show a deeper prediction drop at the stem-suffix boundary than English does (**Figure 5A**). Greek, Italian and Spanish inflection markers are, in fact, more numerous and formally more contrastive than inflection markers are in English.

[12]The non-linear model is statistically more robust, as confirmed by an ANOVA Chi-squared test: p-value < 0.001.

[13]It should be noted that all word forms are centered on the *0*-value, i.e., at the stem-suffix boundary. This makes the interval representing the onset of words less confident, due the possible presence of an inflectional prefix and to differences in length between the stems.

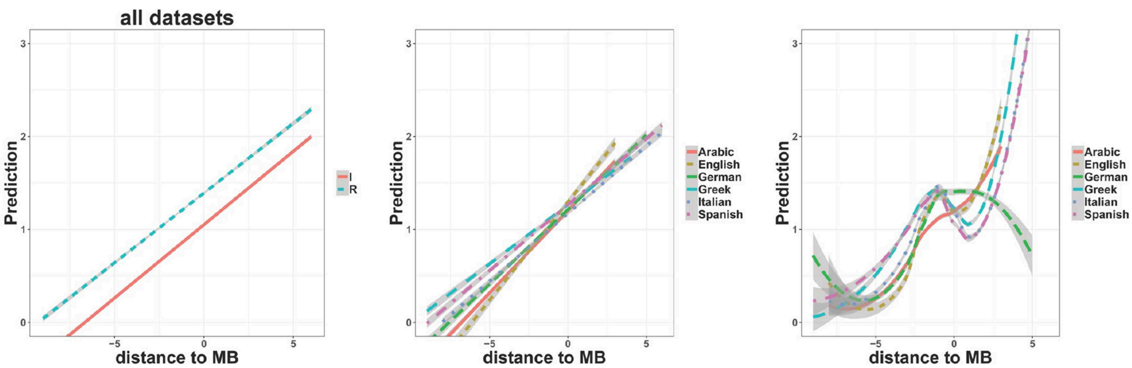


FIGURE 4 | Regression plots of *TSOM* prediction rates in word recoding by letter position to the word stem-suffix boundary (*distance to MB*=0): (possibly prefixed) stems are associated with *x*-values <0, and endings with *x*-values ≥0. **(Left panel)** linear interaction with the categorical regular (cyan) vs. irregular (red) classification of verb paradigms in all languages. **(Center panel)** linear interaction with sample languages as categorical fixed effects. **(Right panel)** non-linear interaction with sample languages as categorical fixed effects. All plots are obtained by using the R *ggplot2* package; related models are respectively calculated with *lm* function and *gam* function (*gamm4* package). Shaded areas indicate 95% confidence intervals.

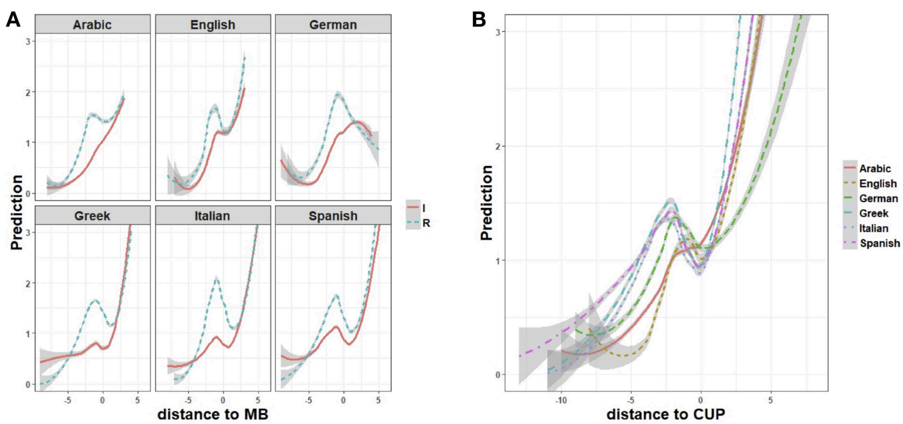


FIGURE 5 | Regression plots (*ggplot2*) of interaction effects **(A)** between regularly (cyan lines) and irregularly (red lines) inflected forms and the distance to stem-suffix boundary (*distance to MB*), and **(B)** between sample languages and the distance to the Complex Uniqueness Point (*distance to CUP*=0), in non-linear models (*GAMs*) fitting the processing prediction rate. Shaded areas indicate 95% confidence intervals.

The steeper ascending trend in the prediction of English forms has to do with the comparative simplicity of the English system, and the narrower range of possible continuations of the verb stem in regularly as well as irregularly inflected forms. The ascending prediction trend of the Arabic verb is more linear, particularly in irregular inflection (**Figure 5A**). In fact, the Arabic plot shows a less prominent drop at the stem-suffix boundary (for *distance to MB*=0), which disappears altogether in irregulars, suggesting a greater structural continuity between the stem and the suffix. This is the combined effect of two factors: stem allomorphy in the perfective forms of irregular verbs, and prefixation of person and number features in the imperfective forms, both concur to reduce processing uncertainty at the stem-suffix boundary[14].

Conversely, German exhibits only descending slopes after the stem-suffix boundary (for *distance to MB*=0). This is an effect of the full formal nesting of shorter endings within longer, onset aligned endings (for instance *e*-, *-en*, *-end*: see **Figure 2** above), which makes it hard to predict upcoming symbols when contextual and frequency information is missing. As a result, prediction increases only at the point where the form is fully discriminated from all other forms (i.e., for *distance to CUP=0*), as shown by the ascending slope for *CUP*>0 in **Figure 5B**.

All in all, closer inspection of prediction curves for each language confirms that regular stems are structurally easier

[14]In Arabic imperfective forms (both regular and irregular), second and third person values are conveyed by the prefixes *ta*- and *ya*- respectively, and number is marked by the suffixes *-u* for the singular, and *-uwna* for the plural. In the first person, person and number are syncretically realized by two different prefixes followed by an identical suffix (*-u*). Hence, the amount of residual processing uncertainty at the stem-suffix boundary is drastically reduced to a binary choice between *-u* and *-uwna* for the second person and the third person only. In perfective forms, regular verbs select one stem followed by different endings for different combinations of number, person and gender features, and irregular verbs select a specific stem allormorph for different combinations of gender and number values with the 3rd person. This further reduces processing uncertainty in the perfective forms of irregular verbs.

to process than irregular stems. Nonetheless, the processing advantage accumulated across regular stems is offset by suffix selection, which requires more effort in adjusting probabilistic expectations for regular forms than for irregular ones. This general, cross-linguistic effect can be interpreted in terms of "surprisal" as a source of processing difficulty (Levy, 2008), caused by the incremental reallocation of the map's expectation for upcoming symbols. In irregular paradigms, the initial processing disadvantage is counterbalanced by a smaller effort (i.e., a smaller decrease in certainty) in processing the transition from a stem to the following suffix. As we will see in more detail in section 5.2.1, this processing facilitation is a consequence of the fact that irregular stems exhibit a formal contrast that regularly inflected forms typically realize with suffixal material. Ultimately, this early contrast provides an early point of lexical discrimination.

5.1.2. Learning Effects

The *TSOM* learning algorithm has a bias to developing specialized chains of processing nodes (*BMUs*), which get more and more sensitive to structural discontinuity in the input data. We can monitor chain specialization by observing how prediction accuracy of online word processing evolves through learning. As training progresses, we expect prediction trends to increasingly reflect the underlying morphological structure of inflected forms. Due to chain specialization, a *TSOM* becomes more and more familiar with recurrent stems and endings. Its ability to predict upcoming symbols increases accordingly, and prediction rates become higher across input words as the map learns to assign maximally entrenched node chains to shared input sub-strings.

Figure 6A plots variation of word prediction rates at different learning epochs (going from epoch 5 to epoch 50 with a 5-epoch step), averaged across all languages in our sample. The prediction rate of inflectional endings starts increasing at epoch 10 (growing slope for *x*-values ≥ 0), when the *TSOM* is not yet able to accurately predict stems (*x*-values <0). This is the combined effect that input frequency and length have on memory. Endings are, on average, more frequent and shorter than stems, and are thus learned more quickly. Processing sensitivity to stems becomes more discriminative only at later epochs, as shown by the local maximum at *distance to MB* $= -1$ at epoch 15. Finally, when stems are predicted accurately (starting from learning epoch 20), the *TSOM* learns that some specific endings follow only some stems, and can adjust its prediction bias accordingly. This makes prediction values for endings (i.e., for *x*-values ≥ 0) increasingly higher at later epochs.

Figure 6B shows the linear growth of processing prediction for each language at different epochs (from epoch 5 to epoch 50). The trend shown by our sample of languages is consistent with the intuition that processing prediction and processing ease increase with learning as a function of memory entrenchment of word representations: the more entrenched word representations are, the more easily they are predicted during processing. This is illustrated by the word-graph in **Figure 2A** (right), which represents the node chains activated by a few forms of German *geben* "give" at an early learning epoch. At this stage, partially overlapping stem allomorphs (*gib-*, *geb-*, and *gab-*) activate superpositional node chains. This indicates that the Markovian order of the map cannot discriminate between *b* preceded by the letter *a*, and the same *b* preceded by the letter *i* or the letter *e*. During training, discriminative specialization results in the word graph being transformed into the tree-like structure of nodes on the left hand side of **Figure 2A**. Note that a word tree has fewer branching points than the corresponding word graph. This means that when the map arrives at a node with fewer branching paths, uncertainty about the upcoming symbol decreases, and processing prediction increases.

Learning, like prediction, is dependent on word length. However, whereas short words are comparatively more difficult to predict, they are learned more easily. Epochs of word learning as a function of word length are compared for each language in **Figure 7A**. Unsurprisingly, short words are learned at early epochs and longer words are learned at later epochs irrespective

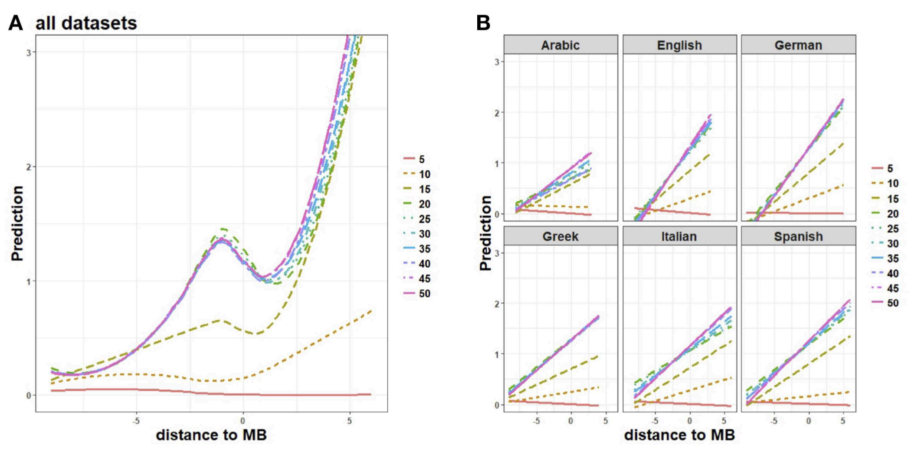


FIGURE 6 | Word structure learning: **(A)** overall *prediction* trends and **(B)** their language-specific linear correlates through 5-to-50 learning epochs, for different letter positions to the stem-suffix boundary (*distance to MB*). Prediction scores exhibit an increasing fit to training data, weighing up the map's expectations across stems (*x*-values <0), and endings (*x*-values $\geq$0).

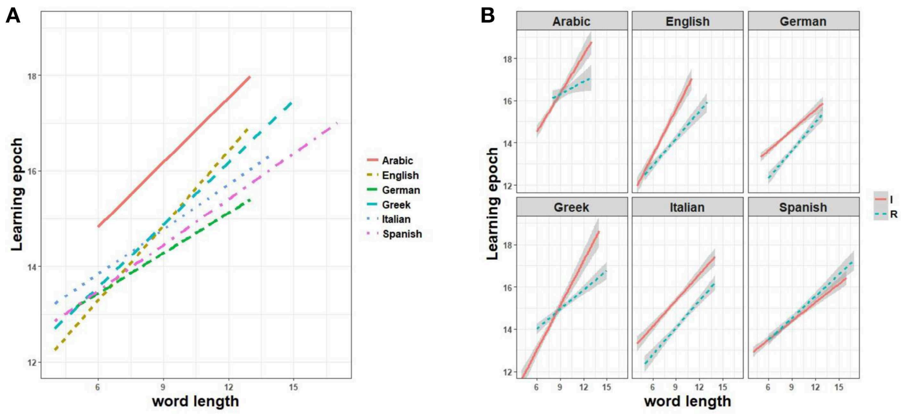


FIGURE 7 | Regression plots of time of lexical acquisition by word length, in linear interaction with **(A)** sample languages and **(B)** regular vs. irregular classification of paradigms plotted for each language.

of language. Nonetheless, the time lag can vary depending on the specific inflection system. For example, English is shown to exhibit a steeper learning slope, i.e., a greater difficulty in learning inflected forms of increasing length, and Arabic shows both a steeper slope and a higher intercept.

When we look at each language in more detail, we observe an interesting interaction between the impact of word length on learning epochs and inflectional regularity. **Figure 7B** plots word learning epochs as a function of word length and the regular vs. irregular dichotomous classification for each language. With the only exception of Spanish (where differences do not reach statistical significance, $p > 0.1$), all languages exhibit different learning paces for regulars vs. irregulars, with the former enjoying advantages over the latter. The pace of learning interacts with word length in Arabic, Greek and English, where long regulars tend to be learned progressively earlier than long irregulars. This contrasts with German, where the trend is reversed, and Italian, which shows no significantly different interaction with increasing length. Note, finally, that when Arabic and Greek irregular forms happen to be significantly shorter than regular forms, the former are learned a few epochs earlier than the latter, reversing the advantage enjoyed by regulars when length is controlled.

5.2. Stem-Family Effects

5.2.1. Processing Effects

The evidence reported in section 5.1 throws into sharp relief a theoretically interesting connection between the structural notions of morphological regularity and transparency, on the one hand, and word processing (un)certainty, on the other hand. *Word-and-Paradigm* approaches to inflection have questioned the primacy of sublexical constituent structure as a key to understanding speaker's word knowledge. In contrast, they support a view of inflectional competence grounded in a complex network of fully-inflected forms organized in word families of lexically-related members (paradigms) and inflectionally-related members (conjugation classes). Accordingly, morphological regularity is the expression of the formal and lexical support that each inflected form receives from its family members: regularly inflected forms tend to get the largest possible amount of formal support from their paradigmatic companions in terms of redundant stem forms (e.g., *walk-s*, *walk-ed*, *walk-ing*), and from their conjugation members in terms of redundant inflectional markers (e.g., *walk-ing*, *see-ing*, *speak-ing*). In this section, we move away from a dichotomous notion of paradigm regularity to a graded one, based on statistical patterns of lexical co-activation and competition. These patterns are better understood, in line with psycholinguistic evidence of human processing behavior, as word family effects.

In section 3.2, we introduced two ways of assessing paradigm regularity as a gradient: one is based on forms, and the other is based on paradigms. Given a target form, we can calculate the number of stem-sharing members of the form's paradigm, or *stem-family size*. For each verb paradigm, the (normalized) average stem-family size of the paradigm gives a graded score of *paradigm regularity*. The two notions are clearly correlated (in our data, $r = 0.79$, $p < 0.001$). In a fully regular paradigm, all forms have the same stem-family size. Conversely, suppletive forms in irregular paradigms tend to be isolated. Nonetheless, the two measures account for different effects on word processing. Intuitively, our paradigm regularity score is the same for all members of a paradigm, irrespective of how stem-families are distributed within the paradigm. However, a form supported by a larger stem-family has a different processing cost from a form with an empty stem-family in the same paradigm. Thus, while paradigm regularity can account for different processing costs between different paradigms, the stem-family size explains the variation, in processing costs, between paradigmatically-related forms.

In our simulations, paradigm regularity has a facilitative effect on stem processing. An identical stem, transparently shared by all paradigmatically-related forms, represents a perceptually salient, deeply entrenched formal core of the paradigm. This core is highly familiar, and is accordingly processed more easily. Conversely, degrees of alternation in stem formation for irregular paradigms tend to slow down the processing of the corresponding stems. The more stem allomorphs a paradigm presents, the larger the effort taken to process them, as more

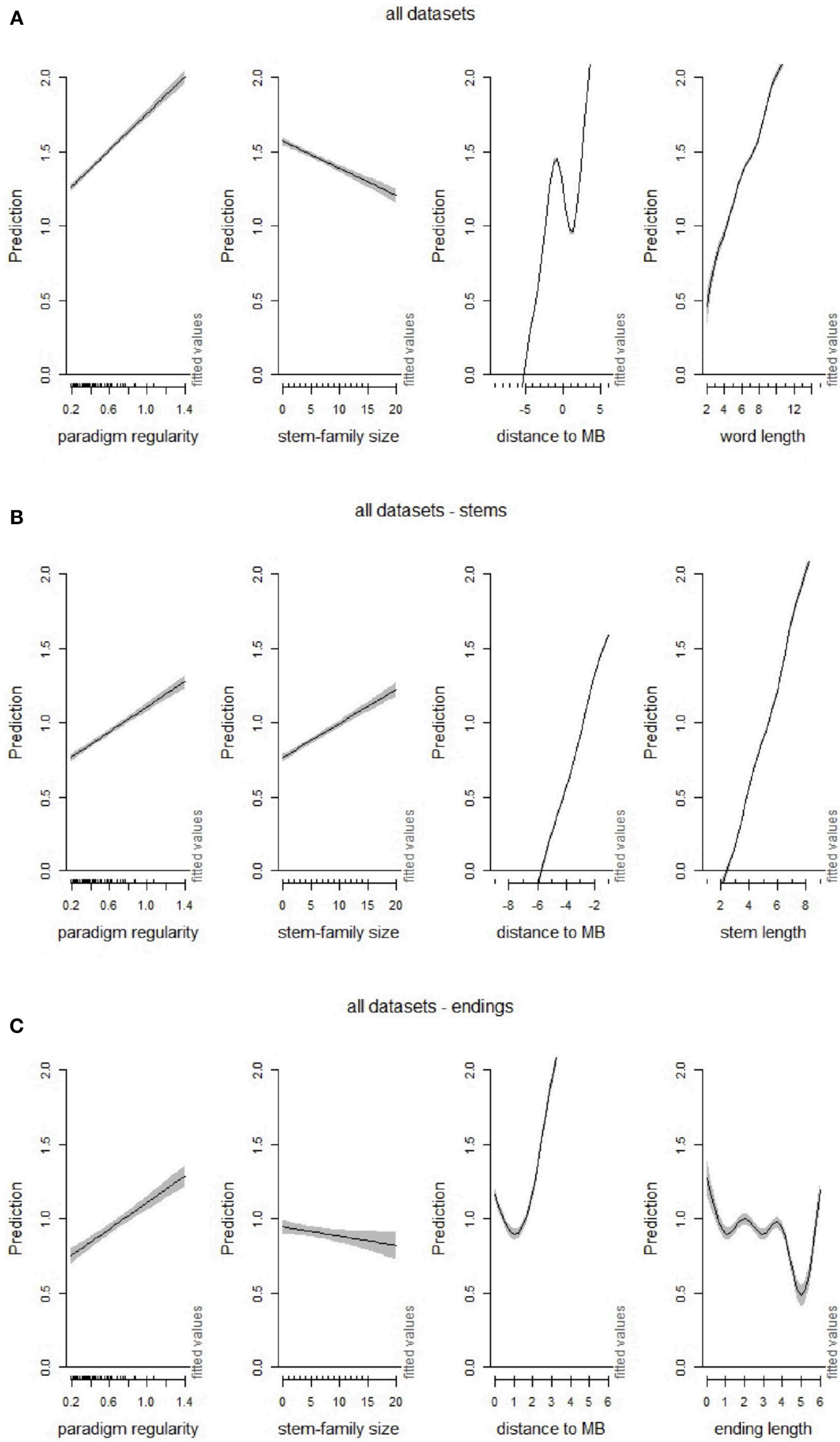


FIGURE 8 | Effects of (graded) paradigm regularity interacting with stem-family size, distance to stem-suffix boundary (*MB*) and (word, stem and suffix) length in *GAM*s modeling *TSOM* prediction rates in: **(A)** word processing, **(B)** stem processing, and **(C)** suffix processing. The specified predictors are taken in interaction with paradigm regularity as summed effects (*itsadug* package, *plot_smooth* function).

points of processing uncertainty are encountered along the way. This is again shown pictorially in **Figure 2A**, where the shaded nodes of the irregular verb *geben*, corresponding to its stem allomorphs, are arranged through branching paths, in contrast with the non-branching path associated with the regular base stem of *glauben* (**Figure 2B**).

Arabic inflection provides a somewhat exemplary illustration of this dynamic. In Arabic, all verb forms present vowel alternating stems, irrespective of their belonging to regular or irregular paradigms. Nonetheless, irregulars are predicted less easily than regulars (**Figure 5**), due to the greater formal diversity that irregular stem allomorphs exhibit. Compare, for example, the regular forms katab-*tu*/katab-*ta*/katab-*a* "I wrote"/"you (singular masculine) wrote"/"he wrote" with the second weak forms qul-*tu*/qul-*ta*/qāl-*a* "I said"/"you (singular masculine) said"/"he said").

The facilitation in processing stems of regularly inflected forms is however counterbalanced by the processing uncertainty incurred at the stem-suffix boundary. In fact, the processing advantage that more regularly inflected forms enjoy across stems is offset by a deeper drop in the prediction rate for an upcoming suffix (**Figure 5**). This is confirmed when we partial out the interacting influence of paradigm regularity, stem-family size, word length and distance to stem-suffix boundary (*MB*) in a generalized additive model of prediction rates (**Figure 8**). Despite the strong positive correlation of paradigm regularity and stem-family size, when these predictors interact, the contributions that they make to the processing of full words go in different directions. When we model the interacting effect of each predictor for stem and ending independently, the facilitation effect of family size on stem processing (**Figure 8B**, second panel from the left) is reversed into a negative effect on endings (**Figure 8C**, second panel from the left).

The interaction of paradigm regularity and stem-family size on word processing is illustrated in more detail through the use of the perspective or contour plots of **Figure 9**, showing the additive interacting effects of paradigm regularity and stem-family size as predictors for processing prediction of full-forms, stems and suffixes. In the plots, contour lines represent prediction values: brown-yellow shades mark areas where items are processed more easily, and blue shades mark areas where items are processed less easily. **Figure 9A** shows that forms in more irregular paradigms are always processed with greater difficulty. However, starting from level 0.4 of paradigm regularity, forms are processed with increasing difficulty if they have large stem-families. Conversely, **Figure 9B** shows that stems appear to be favored by increasing stem-family size, with a mild reversal of this effect only for items in fully regular paradigms (level > 0.8 of paradigm regularity). We can explain these effects by observing that more regular stems are typically followed by a wider range of inflectional endings than highly irregular stems are, and that the number of different endings is proportional to the size of the stem-family. Hence, the larger the stem-family, the higher the processing "surprisal" for upcoming endings (Ferro et al., 2018), as shown in **Figure 9C**.

The consequences of these pervasive stem-family size effects on the perception of morphological structure during word processing are modeled in **Figure 10A**, where non-linear prediction rates are plotted as a function of *distance to MB* for classes of inflected forms with increasing stem-family size (ranging from empty to large). Here, processing curves are aligned on the structural joint between stem and suffix (*distance to MB* $= 0$), confirming that stem predictability increases with the stem-family size, together with the perception of a structural discontinuity at the stem-suffix boundary. In contrast, forms with little or no support from stem-sharing items show lower prediction rates for stems, and higher prediction rates for suffixes. Finally, smaller or no prediction drops at *distance to MB* $= 0$ suggest that an inflected form with an empty-to-small stem-family size is processed and perceived as mono-morphemic.

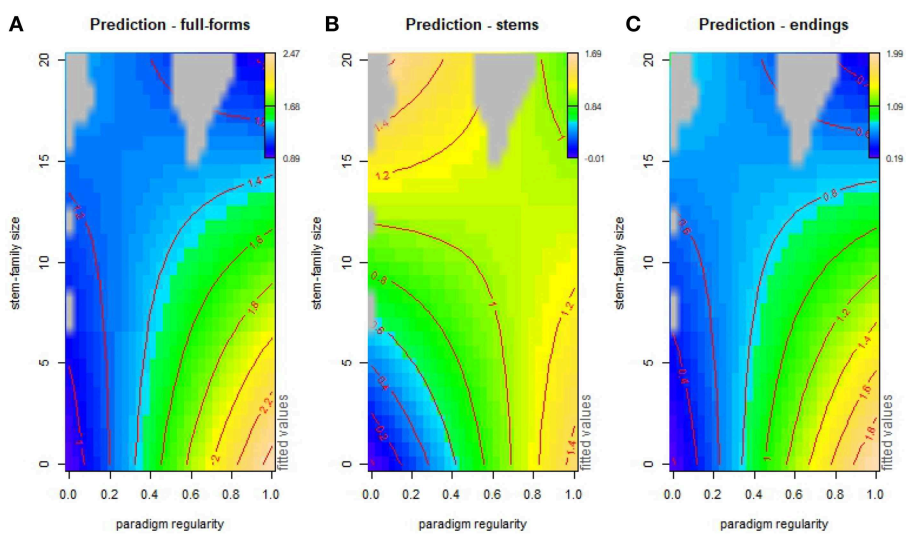


FIGURE 9 | Contour plots of additive interaction effects of stem-family size and paradigm regularity on word **(A)**, stem **(B)**, and ending **(C)** prediction (*itsadug* package, *fvisgam* function). Grid points at a distance of ≥0.2 from the predictor variables are excluded (gray areas).

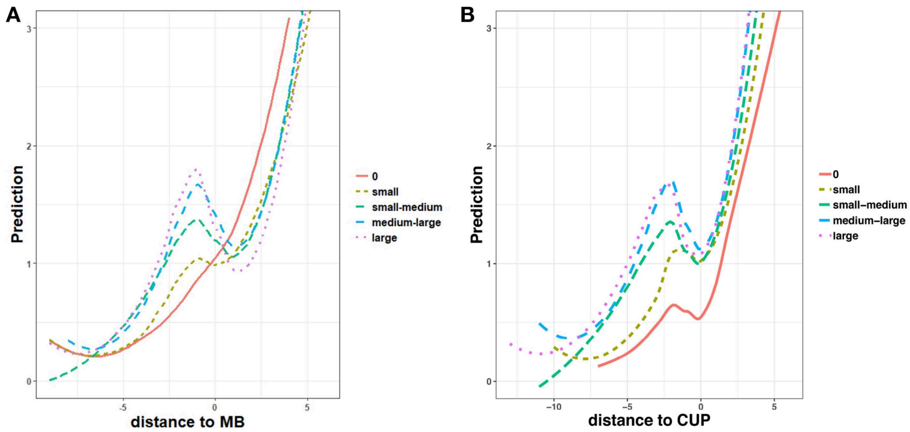


FIGURE 10 | Regression plots of interaction effects between zero, small, small-medium, medium-large and large stem-family sizes and **(A)** distance to stem-suffix boundary (*MB*), and **(B)** distance to *CUP*, in non-linear models (*GAMs*) fitting processing prediction rates.

Another processing effect worth emphasizing in this connection is how easily an inflected form is discriminated from its paradigm members. Discriminability is somewhat related to the predictability of an input sequence. However, whereas predictability is related to the entrenchment of a recurrent sequence in our mental lexicon, discriminability is a function of how confusable an input sequence is with other similar sequences. In fact, full form discrimination requires identification of the form's Complex Uniqueness Point (or *CUP*). Intuitively, the earlier the *CUP* of an inflected form, the more quickly the form is singled out and accessed in serial recognition. Consider two forms like *thought* and *walked*. Unlike *walked*, whose paradigmatic *CUP* is positioned at the beginning of the inflectional ending *-ed*, *thought* is discriminated from all other forms of THINK when *-o* is input. Thus, *thought* is discriminated at an earlier point in time than *walked* is, because of the shorter distance between the form onset and its *CUP* (2 letters, vs. 4 letters for *walked*) and the longer tail of predicted symbols after *CUP* (4 letters, vs. only 1 letter predicted for *walked*). This tendency is confirmed by our analysis of *TSOM* data. Irregulars are, on average, easier to discriminate than regularly inflected items. In **Figure 10B**, we plot non-linear prediction rates as a function of *distance to CUP*, for the same five classes of stem-family size used in **Figure 10A**. For all forms, prediction lines are centered on *CUP* (i.e., for *distance to CUP*=0). Forms with empty stem families show a shorter time lag between their onset and their *CUP*. In addition, they have a longer tail of predicted symbols after *CUP*. As we move from isolated forms to forms with increasing stem-family size, the time interval between onset and *CUP* increases, and the tail of symbols after *CUP* shortens. Family size is a significant predictor of this trend, even when the length of inflected forms is added in interaction to the *GAM*.

To sum up, stem-family size appears to influence the rate at which *TSOM*s process verb forms, confirming that paradigm regularity facilitates processing. The stems of inflected forms with a large stem-family size are easier to predict, but this facilitation is compensated at the stem-suffix boundary, where a significant reduction in processing prediction can be interpreted as a functional correlate of structural discontinuity. Conversely, the stems of inflected items with empty-to-small stem-family size are more difficult to predict, and take more processing effort, but show a considerably smaller drop in prediction at *MB*, which suggests that they are processed more holistically. In addition, irregulars are discriminated more quickly from their paradigmatic companions than regularly inflected forms are. It is noteworthy that this effect is not an artifact of irregulars being, on average, shorter than regulars, but reflects a genuine morphological pattern: irregulars tend to mark, through stem allomorphy, information that regularly inflected forms mark with their suffixes. Ultimately, irregularly inflected forms are easier to discriminate because they are part of a more contrastive network of formal oppositions.

5.2.2. Learning Effects

From a learning perspective, being part of a larger stem-family is an advantage. Intuitively, the size of a stem-family defines the number of distinct affixes the stem can combine with: the larger the stem-family size, the greater the potential of the stem for filling in more paradigm cells. Hence, a stem-family defines the analogical space where a verb establishes its connections with other paradigms. Other things being equal (e.g., length, cumulative frequency, and wordlikeness), a verb form with a large stem-family is learned more easily than an isolated form, as shown in **Figure 11A**, where we plot a largely predominant facilitation effect of stem-family size on word learning across the entire length range. A regular paradigm is a classical example of such a stem-family, where all family members share the same stem followed by a systematic pool of inflectional endings (which are, in turn, shared by other paradigms). Being associated with a large range of differently inflected forms makes it possible for a *TSOM* to consolidate intra-paradigmatic formal redundancies, and infer missing forms more accurately. Conversely, more discriminated irregular forms are acquired more slowly for exactly the same reason that they are

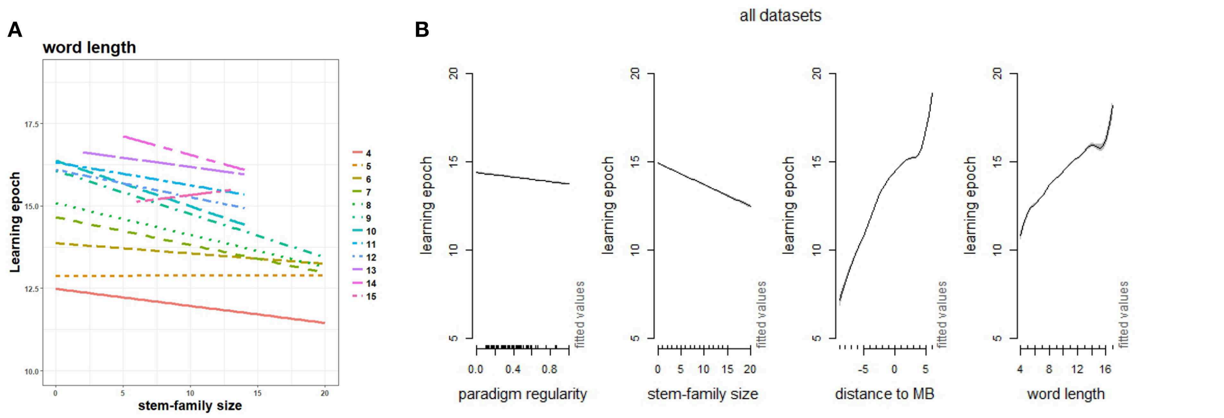


FIGURE 11 | **(A)** Regression plot of time of lexical acquisition by size of stem-family in linear interaction with word length. Fixed effects are word length (as categorical predictor), and stem-family size; **(B)** summed effects of paradigm regularity interacting with stem-family size, distance to stem-suffix boundary (*MB*) and word length in a *GAM* modeling time of acquisition (learning epoch) for verb forms in all sample languages.

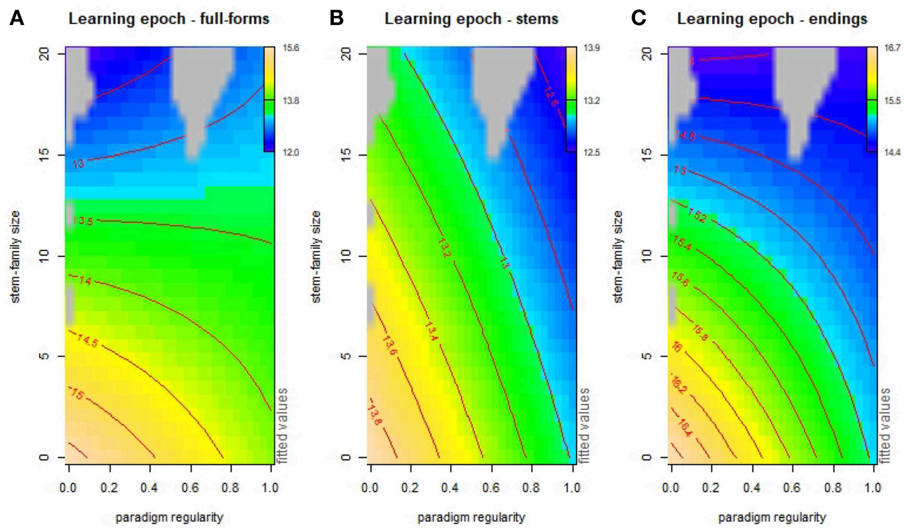


FIGURE 12 | Contour plots of additive interaction effects of stem-family size and paradigm regularity on word **(A)**, stem **(B)**, and ending **(C)** learning. Grid points at a distance of ≥0.2 from the predictor variables are excluded (gray areas).

accessed more holistically and effectively: because they happen to be isolated.

In **Figure 11B**, we assess the interacting influence of paradigm regularity, stem-family size, word length and *distance to MB* in a generalized additive model of the pace of word acquisition. In this model, acquisition is jointly facilitated by both factors, unlike what we observed for processing prediction (**Figure 8A**) where paradigm regularity and stem-family size have contrasting effects. The contour plots of **Figure 12** show the interaction of these factors in more detail. Stem-family is shown to have a main facilitation effect on word learning, which is stronger within weakly regular paradigms. This suggests that even when a verb is associated with many unpredictable stems, the distribution of these stems in their families affects learning. Regular paradigms are facilitative too, but their influence is comparatively less prominent. It is noteworthy that this general word-level effect is differently apportioned when we focus on the pace of stem learning: stems are learned significantly more quickly within regular paradigms, whereas the size of their stem-family has only a marginal (positive) effect in the process. Finally, endings are learned more quickly in regular paradigms, but the facilitative influence of the size of the stem-family in this case is considerably more significant.

Finally, to assess the independent impact of stem-family size on word learning, we ran a generalized additive model including stem-family size, word length and distance to stem-suffix boundary as predictors for learning epoch (R^2(adj) = 0.53), and

TABLE 2 | *GAM* fitted to word learning epochs using stem-family size as the only paradigmatic predictor: *learning epoch∼ stem-family size*distance to MB*word length* + s(*distance to MB*) + s(*word length*).

Parametric coefficients	Estimate	Std. Error	*t* value	Pr(>t)
(Intercept)	0.1677863	0.0648875	2.586	0.00972
Distance to MB	4.7535561	0.1025123	46.371	< 2e-16
Stem-family size	−0.1873784	0.0105171	−17.816	< 2e-16
Word length	1.9595085	0.0188123	104.161	< 2e-16
Distance to MB:stem-family size	−0.0553514	0.0031115	−17.789	< 2e-16
Distance to MB:word length	−0.1973904	0.0031048	−63.577	< 2e-16
Stem-family size:word length	0.0047230	0.0011342	4.164	3.13e-05
Distance to MB:stem-family size:word length	0.0060111	0.0003114	19.305	< 2e-16
Significance of smooth terms:				
s(distance to MB)				< 2e-16
s(word length)				< 2e-16
R^2(adj) = 0.53				

compared it with a similar model where the paradigmatic predictor is paradigm regularity (R^2(adj) = 0.28). The robustness of the stem-family model (see **Table 2**) provides solid support for the hypothesis that family size is the driving paradigmatic force in word learning[15].

6. GENERAL DISCUSSION AND CONCLUDING REMARKS

Inflectional complexity is inherently multidimensional. It lends itself rather reluctantly to an assessment in terms of a single factor or measure. Traditional approaches to complexity are based on either an inventory of morphosyntactic features and their markers, or on a full grammatical description of an inflection system. Both strategies require considerable preliminary knowledge about lexical/grammatical units as well as rules/processes for their segmentation and recombination. Ultimately, they appear to place more emphasis on the descriptive adequacy of some system of formalized knowledge, than on inflectional complexity *per se*.

More recently, information-theoretic approaches have tried to eschew such a circular assessment of system complexity, by looking at patterns of interpredictability in the distribution of inflectional data. These patterns are not scattered randomly across inflectional paradigms. On the contrary, they are distributed in ways that reduce the amount of uncertainty in mastering inflection. Such a structured system of implicative relations addresses a few learnability issues by constraining an otherwise very large hypothesis space. This is certainly a step forward in understanding inflection and measuring its complexity. However, there are a number of important issues that it leaves unaddressed.

Form interpredictability requires preliminary identification of patterns of morphological redundancy, such as those allowing the resolution of a simple analogical proportion like <*walk* :: *walked* = *talk* :: ?>. However, the hypothesis space of formal mapping may vary considerably from language to language, as witnessed by proportions like <*machen* :: *gemacht* = *lachen* :: ?> (German) and <*kataba* :: *yaktubu* = *hadama* :: ?> (Arabic). For the idea of measuring the complexity of an inflectional system in terms of information entropy to be empirically verifiable, we have to be fairly specific about how speakers can acquire these patterns in the first place, and what learning principles provide the algorithmic basis for bootstrapping patterns with no *a priori* knowledge about them. What cues allow the identification of sublexical parts? Do these parts provide, in turn, functional cues to lexical processing? Are all morphological systems equally complex to process and learn? And if they are not, why?

To meet its communicative goals any inflection system needs to be *learnable* in the first place. In addition, it must also be *functional* from the standpoint of processing requirements, and its inner organization must address fundamental discriminative concerns about efficiency in lexical access. In particular, inflected forms should not simply be easy to generalize, but also easy to process and amenable to accurate and efficient access. Issues of learnability, processability and discriminability are not necessarily mutually contradictory, but they may pull in different directions. Examining their complex interaction requires much more than just a way to gauge inferential entropy in a static repository of inflected forms. In this paper, we suggested using basic discriminative learning principles and a type of recurrent artificial network to clarify the multidimensional nature of inflectional complexity in the light of our understanding of the interaction between learning principles and human word processing behavior.

As a general point, the evidence offered in this paper shows that the processing cost of typologically different inflectional systems oscillates within a fairly narrow range (**Figure 4**, center panel), and that this range is actually modulated by morphological structure (**Figure 4**, right panel). It is noteworthy that the steepest linear slopes of processing prediction in the center panel of **Figure 4** are associated with Arabic and

[15]In **Table 2**, predicted, standard error and *p* values are detailed for each of the independent variables. The adjusted *R-squared* expresses the percentage of explained deviance. The intercept value is referred to our dependent variable (learning epoch) when all the independent predictors are 0.

English, while German, Greek, Italian and Spanish present less steep word processing trends. It would be an unwarranted oversimplification, however, to conclude that Arabic and English have *less* complex inflection systems. In Arabic, the effect is mainly due to the discontinuous structure of verb stem allomorphs, which systematically anticipate the marking of a few morphosyntactic features in both regulars and irregulars, and make endings easier to predict. The result is a lower rate of processing uncertainty at the stem-suffix boundary, compared with other languages. A similar processing facilitation is observed in English for a different reason: the small range of endings marking inflectional contrast. Conversely, German, Greek, Italian and Spanish show higher prediction rates for stem processing, followed by a deeper drop in prediction at the stem-suffix boundary, due to a richer set of (possibly nested) inflectional suffixes. Languages typologically vary in the ways they apportion morpho-syntactic information across inflected words. This variation has a significant impact on word processing strategies, because finding information at certain points in the input word may reduce processing uncertainty at later points. Nonetheless, when the amount of morpho-syntactic information to convey through inflection is approximately constant across languages[16], cross-linguistic variation in processing uncertainty can only oscillate within a narrow range.

This has also interesting repercussions on learning. Judging from the distribution of per-word learning epochs (**Figure 3A**), Arabic is arguably the most difficult language to learn in our sample. In *TSOM*s, this is caused by the strong paradigmatic competition between discontinuous roots intercalated by different vowel patterns. Nonetheless, the system appears to substantially benefit from interpredictability patterns, as shown by the distribution of learning span values in **Figure 3B**. This suggests that the higher cost of learning discontinuous stems in Arabic is compensated by the relatively shorter length of inflected forms (compared with the length of inflected forms in other languages), and by the fact that inflectional class information is conveyed by the stem. In other languages like Greek, Italian and Spanish, where class information is mostly conveyed after the stem, selection of the appropriate inflectional ending is considerably more uncertain (in line with Ackerman and Malouf's LCEC, Ackerman and Malouf, 2013).

From a functional perspective, the simulation evidence offered in this paper can be interpreted as the result of a balancing act between two potentially competing communicative requirements: (i) a recognition-driven tendency for a maximally contrastive system, where all inflected forms, both within and across paradigms, present the earliest possible (complex) uniqueness point of recognition; and (ii) a production-driven bias for a maximally generalizable inflection system, where most forms in a paradigm can be inferred from any other form in the same paradigm. In an efficient communication system, both requirements should be balanced. A poorly contrastive system would be dysfunctional for practical communication purposes, because the elements in the system are harder to discriminate. A communication system, however, must also be open and adaptive, because new elements must continuously be added to be able to refer to an ever changing pragmatic environment. Clearly, a maximally contrastive system of irregular forms would take the least effort to process, but would require full storage of unpredictable items, thus turning out to be slow to learn. A maximally generalizable system, on the other hand, would be comparatively easier to learn, but rather inefficient to process, especially when it comes to low-frequency items.

To conclude, there are a few implications of the present work that are worth emphasizing here because of their impact on our current understanding of word knowledge. First, from observing discriminative learning principles in action, we get a very concrete sense of the view that processing and learning issues are strongly inter-related and mutually-implied in language. Although they may pull in different directions, they are, in fact, two sides of the same coin. More precisely, they define two distinct temporal dynamics (a short-term one for processing, and a long-term one for learning) of the same unitary underlying process. Accordingly, representations of word forms in our mental lexicon are only stored patterns of their processing history: what we learn of language is basically how we process it (Marzi and Pirrelli, 2015). Secondly, temporal patterns of discontinuous redundant units in time-series are known to be more difficult to process by humans (Hahn and Bailey, 2005). Hence, it is expected that they should be more difficult to learn and organize in our mental lexicon. This point has interesting implications both for our categorization of inflectional processes into regular and irregular, and in cross-linguistic terms. Systems of formally alternating complex units are easier to discriminate, since they provide shorter *CUPs* and thus take, overall, a smaller processing surprisal (Balling and Baayen, 2008, 2012). However, they are more difficult to acquire, since they are more isolated from other existing forms, and thus harder to infer from familiar, redundant patterns. Less discriminative, highly syncretic systems are much easier to generalize and learn, and, for the same reason, more ambiguous (less informative) and more difficult to interpret in their context of use. Such a complex dynamic of co-existing and conflicting processing requirements has to find a balance, both within the same language (as a function of how regulars and irregular items are stored and co-organized in the mental lexicon), and cross-linguistically (as a function of typologically diverse processes of inflection marking). In the end, each language (and arguably each individual learner) is likely to strike a different balance, which nonetheless falls within a reasonably tight range of variation, bounded by a few learnability and processability requirements. This suggests that full investigation of morphological systems will likely benefit

[16]We are not suggesting that such an invariance in the amount of inflectional information holds for all languages. For example, Basque verb agreement marks an inflected verb form with affixes for subject, direct object and indirect object case (Austin, 2010, 2012). The system is agglutinative, and the number of possible distinct affix combinations for ditransitive verbs soon gets very large (up to 102 different forms in the present indicative of the auxiliary). Thus the amount of syntactic context that must be processed for a child to check case assignment on the main verb form, is considerably larger than what is needed for languages like German, Italian or Spanish.

from the use of basic concepts from the toolkit of complexity theory in biological networks, such as emergence, non-linearity and self-organization.

Over the past few decades, morphological paradigms have acquired growing importance in our understanding of complex inflectional systems. While our work provides novel evidence supporting this trend, it also suggests that paradigms are functionally specialized by-products of more general principles of lexical self-organization, based on the more primitive notion of word neighborhood. The morphological correlate of this notion, namely the family of inflected types sharing the same stem, or stem-family, turned out to be key to understanding processing and learning issues. The size and distributional entropy of stem-families can ultimately account for the effects of ease of generalization, thus providing a more general, process-oriented basis for understanding patterns of interpredictability in morphology learning. This perspective lends strong support to the view that the complexity of the inflectional system of a language is a collective, emergent property that cannot be deduced from its individual forms. In addition, it suggests that a processing-oriented, discriminative view of inflection systems is not incompatible with the idea that their systemic self-organization rests on smaller "morphomic" patterns of structural redundancy.

AUTHOR CONTRIBUTIONS

CM ran the experiments on German, English, Arabic and Spanish, conducted data analysis and statistical modeling of the results, and developed the conceptual and theoretical framework of this work. MF was responsible for computational modeling, ran the experiments on Italian and Greek, designed and extracted functional responses and structural evidence of TSOMs for quantitative data analysis. VP developed the linguistic framework and the theoretical background of the present study. Theoretical implications and concluding remarks were jointly discussed. CM and VP critically revised previous versions of the paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2019.00048/full#supplementary-material

REFERENCES

Ackerman, F., Blevins, J. P., and Malouf, R. (2009). "Parts and wholes: patterns of relatedness in complex morphological systems and why they matter," in *Analogy in Grammar: Form and Acquisition*, eds J. P. Blevins and J. Blevins (Oxofrd, UK: Oxford University Press), 54–82.

Ackerman, F., and Malouf, R. (2013). Morphological organization: the low conditional entropy conjecture. *Language* 89, 429–464. doi: 10.1353/lan.2013.0054

Albright, A. (2002). Islands of reliability for regular morphology: evidence from italian. *Language* 78, 684–709. doi: 10.1353/lan.2003.0002

Althaus, N., and Mareschal, D. (2013). Modeling cross-modal interactions in early word learning. *IEEE Trans. Auton. Ment. Dev*. 5, 288–297. doi: 10.1109/TAMD.2013.2264858

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: resolving neighborhood conflicts. *Psychon. Bull. Rev*. 4, 439–461. doi: 10.3758/BF03214334

Aronoff, M. (1994). *Morphology by Itself: Stems and Inflectional Classes*. Number 22. Cambridge, MA: The MIT Press. doi: 10.2307/416331

Austin, J. (2010). Rich inflection and the production of finite verbs in child language. *Morphology* 20, 41–69. doi: 10.1007/s11525-009-9144-7

Austin, J. (2012). The case-agreement hierarchy in acquisition: evidence from children learning basque. *Lingua* 122, 289–302. doi: 10.1016/j.lingua.2011.10.013

Baayen, H. R., Piepenbrock, P., and Gulikers, L. (1995). *The CELEX Lexical Database (CD-ROM)*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania. p. 438–481.

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., and Blevins, J. P. (2019). The discriminative lexicon: a unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de) composition but in linear discriminative learning. *Complexity* 2019:4895891. doi: 10.1155/2019/4895891

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychol. Rev*. 118, 438–481. doi: 10.1037/a0023851

Balling, L. W., and Baayen, R. H. (2008). Morphological effects in auditory word recognition: evidence from danish. *Lang. Cogn. Process.* 23, 1159–1190. doi: 10.1080/01690960802201010

Balling, L. W., and Baayen, R. H. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition* 125, 80–106. doi: 10.1016/j.cognition.2012.06.003

Bane, M. (2008). "Quantifying and measuring morphological complexity," in *Proceedings of the 26th West Coast Conference on Formal Linguistics* (Berkeley, CA: University of California), 69–76.

Bates, E., and MacWhinney, B. (1987). "Competition, variation, and language learning," in *Mechanisms of Language Acquisition*, ed B. MacWhinney (Hillsdale, NJ: Lawrence Erlbaum Associates), 157–193.

Bearman, M., Brown, D., and Corbett, G. G., editors (2015). *Understanding and Measuring Morphological Complexity*. Oxford, UK: Oxford University Press.

Bickel, B., and Nichols, J. (2005). "Inflectional synthesis of the verb," in *The World Atlas of Language Structures*, eds M. Haspelmath, M. S. Dryer, D. Gil, and B. Comrie (Oxford, UK: Oxford University Press), 94–97.

Bittner, D., Dressler, W. U., and Kilani-Schoch, M., editors (2003). *Development of Verb Inflection in First Language Acquisition: A Cross-Linguistic Perspective*. Berlin: Mouton de Gruyter.

Blevins, J. P. (2016). *Word and Paradigm Morphology*. Oxford, UK: Oxford University Press.

Blevins, J. P., Milin, P., and Ramscar, M. (2017). "The zipfian paradigm cell filling problem," in *Morphological Paradigms and Functions*, eds F. Kiefer, J. P. Blevins, and H. Bartos (Leiden: Brill), 141–158.

Blom, E. (2007). Modality, infinitives, and finite bare verbs in Dutch and English child language. *Lang. Acquisit.* 14, 75–113. doi: 10.1080/10489220701331821

Bompolas, S., Ferro, M., Marzi, C., Cardillo, F. A., and Pirrelli, V. (2017). For a performance-oriented notion of regularity in inflection: the case of modern greek conjugation. *Ital. J. Comput. Linguist.* 3, 77–92. doi: 10.4000/books.aaccademia.1721

Bonami, O., and Beniamine, S. (2016). Joint predictiveness in inflectional paradigms. *Word Struct*. 9, 156–182. doi: 10.3366/word.2016.0092

Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S.-Y., Pascual, L., et al. (2004). Cross-linguistic analysis of vocabulary in young children: Spanish, dutch, french, hebrew, italian, korean, and american english. *Child Dev*. 75, 1115–1139. doi: 10.1111/j.1467-8624.2004.00729.x

Chen, Q., and Mirman, D. (2012). Competition and cooperation among similar representations: toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychol. Rev.* 119:417. doi: 10.1037/a0027175

Crago, M. B., and Allen, S. E. (2001). Early finiteness in inuktitut: the role of language structure and input. *Lang. Acquisit.* 9, 59–111. doi: 10.1207/S15327817LA0901_02

Dimitropoulou, M., Duñabeitia, J. A., Avilés, A., Corral, J., and Carreiras, M. (2010). Subtitle-based word frequencies as the best estimate of reading behavior: the case of greek. *Front. Psychol.* 1:218. doi: 10.3389/fpsyg.2010.00218

Elman, J. L. (1990). Finding structure in time. *Cogn. Sci.* 14, 179–211. doi: 10.1207/s15516709cog1402_1

Evans, N., and Levinson, S. C. (2009). The myth of language universals: language diversity and its importance for cognitive science. *Behav. Brain Sci.* 32, 429–492. doi: 10.1017/S0140525X0999094X

Ferro, M., Marzi, C., and Pirrelli, V. (2011). A self-organizing model of word storage and processing: implications for morphology learning. *Lingue e Linguaggio* 10, 209–226. doi: 10.1418/35840

Ferro, M., Marzi, C., and Pirrelli, V. (2018). Discriminative word learning is sensitive to inflectional entropy. *Lingue e linguaggio* 17, 307–327. doi: 10.1418/91871

Finkel, R., and Stump, G. (2007). Principal parts and morphological typology. *Morphology* 17, 39–75. doi: 10.1007/s11525-007-9115-9

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Comput. Linguist.* 27, 153–198. doi: 10.1162/089120101750300490

Goodman, J. C., Dale, P. S., and Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *J. Child Lang.* 35, 515–531. doi: 10.1017/S0305000907008641

Hahn, U., and Bailey, T. M. (2005). What makes words sound similar? *Cognition* 97, 227–267. doi: 10.1016/j.cognition.2004.09.006

Haspelmath, M., Dryer, M. S., Gil, D., and Comrie, B., (eds.). (2005). *The World Atlas of Language Structures*, Vol. 1. Oxford, UK: Oxford University Press.

Juola, P. (1998). Measuring linguistic complexity: the morphological tier. *J. Quant. Linguist.* 3, 206–213. doi: 10.1080/09296179808590128

Kohonen, T. (2002). *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Berlin: Springer.

Kolmogorov, A. N. (1968). Three approaches to the quantitative definition of information. *Int. J. Comput. Math.* 2, 157–168. doi: 10.1080/00207166808803030

Krajewski, G., Lieven, E. V., and Theakston, A. L. (2012). Productivity of a polish child's inflectional noun morphology: a naturalistic study. *Morphology* 22, 9–34. doi: 10.1007/s11525-011-9199-0

Kuperman, V., Bertram, R., and Baayen, R. H. (2010). Processing trade-offs in the reading of dutch derived words. *J. Mem. Lang.* 62, 83–97. doi: 10.1016/j.jml.2009.10.001

Legate, J. A., and Yang, C. (2007). Morphosyntactic learning and the development of tense. *Lang. Acquisit.* 14, 315–344. doi: 10.1080/10489220701471081

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition* 106, 1126–1177. doi: 10.1016/j.cognition.2007.05.006

Li, P., Zhao, X., and Mac Whinney, B. (2007). Dynamic self-organization and early lexical development in children. *Cogn. Sci.* 31, 581–612. doi: 10.1080/15326900701399905

Luce, P. A., and Pisoni, D. B. (1998). Recognizing spoken words: the neighborhood activation model. *Ear Hear.* 19:1. doi: 10.1097/00003446-199802000-00001

Lyding, V., Stemle, E., Borghetti, C., Brunello, M., Castagnoli, S., Dell' Orletta, F., et al. (2014). "The paisá corpus of italian web texts," in *Proceedings of the 9th Web as Corpus Workshop (WaC-9)@ EACL 2014* (Gothenburg: Association for Computational Linguistics), 36–43.

Maamouri, M., Bies, A., Buckwalter, T., and Mekki, W. (2004). "The penn arabic treebank: building a large-scale annotated arabic corpus," in *NEMLAR Conference on Arabic Language Resources and Tools* (Paris: ELDA), Vol. 27, 466–467.

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., and Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cogn. Sci.* 31, 133–156. doi: 10.1080/03640210709336987

Marslen-Wilson, W. D. (1984). "Function and process in spoken word recognition: a tutorial review," in *Attention and Performance X: Control of Language Processes*, eds H. Bouma and D. G. Bouwhuis (New Jersey, NJ: Lawrence Erlbaum Associates), 125–150.

Marzi, C., Ferro, M., and Nahli, O. (2017). Arabic word processing and morphology induction through adaptive memory self-organisation strategies. *J. King Saud Univ. Comput. Inform. Sci.* 29, 179–188. doi: 10.1016/j.jksuci.2016.11.006

Marzi, C., Ferro, M., Nahli, O., Belik, P., Bompolas, S., and Pirrelli, V. (2018). "Evaluating inflectional complexity crosslinguistically: a processing perspective," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan*, eds N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga (Paris: European Language Resources Association (ELRA)), 3860–3866.

Marzi, C., Ferro, M., and Pirrelli, V. (2012). Word alignment and paradigm induction. *Lingue e Linguaggio* XI, 251–274.

Marzi, C., Ferro, M., and Pirrelli, V. (2014). Morphological structure through lexical parsability. *Lingue e Linguaggio* 13, 263–290.

Marzi, C., and Pirrelli, V. (2015). A neuro-computational approach to understanding the mental lexicon. *J. Cogn. Sci.* 16, 493–535. doi: 10.17791/jcs.2015.16.4.492

Matthews, P. H. (1991). *Morphology*. Cambridge: Cambridge University Press.

Mayor, J., and Plunkett, K. (2010). A neurocomputational account of taxonomic responding and fast mapping in early word learning. *Psychol. Rev.* 117:1. doi: 10.1037/a0018130

McWorther, J. (2001). The world's simplest grammars are creole grammars. *Linguist. Typol.* 5, 125–166. doi: 10.1515/lity.2001.001

Milin, P., Đurđević, D. F., and del Prado Martín, F. M. (2009b). The simultaneous effects of inflectional paradigms and classes on lexical recognition: evidence from serbian. *J. Mem. Lang.* 60, 50–64. doi: 10.1016/j.jml.2008.08.007

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., and Baayen, R. H. (2017). Discrimination in lexical decision. *PLoS ONE* 12:e0171935. doi: 10.1371/journal.pone.0171935

Milin, P., Kuperman, V., Kostic, A., and Baayen, R. H. (2009a). "Paradigms bit by bit: an information theoretic approach to the processing of paradigmatic structure in inflection and derivation," in *Analogy in Grammar: Form and Acquisition*, eds J. P. Blevins and J. Blevins (Oxford, UK: Oxford University Press), 241–252.

Moscoso del Prado Martín, F., Kostić, A., and Baayen, R. H. (2004). Putting the bits together: an information theoretical perspective on morphological processing. *Cognition* 94, 1–18. doi: 10.1016/j.cognition.2003.10.015

Phillips, C. (1996). "Root infinitives are finite," in *Proceedings of the 20th Annual Boston University Conference on Language Development* (Boston, MA: Cascadilla Press), 588–599.

Phillips, C. (2010). Syntax at age two: cross-linguistic differences. *Lang. Acquisit.* 17, 70–120. doi: 10.1080/10489221003621167

Pine, J. M., Conti-Ramsden, G., Joseph, K. L., Lieven, E. V., and Serratrice, L. (2008). Tense over time: testing the agreement/tense omission model as an account of the pattern of tense-marking provision in early child english. *J. Child Lang.* 35, 55–75. doi: 10.1017/S0305000907008252

Pirrelli, V. (2000). *Paradigmi in Morfologia. Un Approccio Interdisciplinare Alla Flessione Verbale Dell'Italiano*. Pisa: Istituti Editoriali e Poligrafici Internazionali.

Pirrelli, V., and Battista, M. (2000). The paradigmatic dimension of stem allomorphy in italian verb inflection. *Ital. J. Linguist.* 12, 307–380.

Pirrelli, V., Ferro, M., and Marzi, C. (2015). "Computational complexity of abstractive morphology," in *Understanding and Measuring Morphological Complexity*, eds M. Baerman, D. Brown, and G. G. Corbett (Oxford, UK: Oxford University Press), 141–166.

Pirrelli, V., and Yvon, F. (1999). The hidden dimension: a paradigmatic view of data-driven nlp. *J. Exp. Theor. Artif. Intell.* 11, 391–408. doi: 10.1080/095281399146472

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Ralli, A. (2005). *Morfologia [Morphology]*. Athens: Patakis.

Ralli, A. (2006). *On the Role of Allomorphy in Inflectional Morphology: Evidence From Dialectal Variation* Monza: Polimetrica.

Ralli, A. (2014). "Suppletion," in *Encyclopedia of Ancient Greek language and linguistics*, ed G. K. Giannakis (Leiden: Brill), 341–344.

Ramscar, M., and Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: a new approach to the logical problem of language acquisition. *Cogn. Sci.* 31, 927–960. doi: 10.1080/03640210701703576

Rescorla, R. A., and Wagner, A. R. (1972). "A theory of pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement," in *Classical Conditioning II: Current Research and Theory*, eds A. H. Black and W. F. Prokasy (New York, NY: Appleton-Century-Crofts), 64–99.

Rissanen, J. (2007). *Information and Complexity in Statistical Modeling*. Berlin: Springer Science & Business Media.

Sagot, B. (2018). *Informatiser le lexique: Modélisation, développement et utilisation de lexiques morphologiques, syntaxiques et sémantiques*. (Thèse d'Habilitation). Université Paris-Sorbonne, Paris, France.

Sagot, B., and Walther, G. (2011). "Non-canonical inflection: data, formalisation and complexity measures," in *International Workshop on Systems and Frameworks for Computational Morphology* (Berlin: Springer), 23–45.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Shosted, R. (2006). Correlating complexity: a typological approach. *Linguist. Typol.* 10, 1–40. doi: 10.1515/LINGTY.2006.001

Slobin, D. I. (1982). "Universal and particular in the acquisition of language," in *Language Acquisition: The State of the Art*, eds L. R. Gleitman and E. Wanner (Cambridge, UK: Cambridge University Press), 128–172.

Slobin, D. I., editor (1985). *The Crosslinguistic Study of Language Acquisition*, Vol. 1. New Jersey, NJ: Lawrence Erlbaum Associates.

Wilson, S. (2003). Lexically specific constructions in the acquisition of inflection in english. *J. Child Lang.* 30, 75–115. doi: 10.1017/S0305000902005512

Xanthos, A., Laaha, S., Gillis, S., Stephany, U., Aksu-Koç, A., Christofidou, A., et al. (2011). On the role of morphological richness in the early development of noun and verb inflection. *First Lang.* 31, 461–479. doi: 10.1177/0142723711409976

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 10 October 2019
doi: 10.3389/fpsyg.2019.02162

Understanding Events by Eye and Ear: Agent and Verb Drive Non-anticipatory Eye Movements in Dynamic Scenes

Roberto G. de Almeida[1]*, *Julia Di Nardo*[1], *Caitlyn Antal*[1,2] *and Michael W. von Grünau*[1†]

[1] Department of Psychology, Concordia University, Montreal, QC, Canada, [2] Department of Linguistics, Yale University, New Haven, CT, United States

As Macnamara (1978) once asked, how can we talk about what we see? We report on a study manipulating realistic dynamic scenes and sentences aiming to understand the interaction between linguistic and visual representations in real-world situations. Specifically, we monitored participants' eye movements as they watched video clips of everyday scenes while listening to sentences describing these scenes. We manipulated two main variables. The first was the semantic class of the verb in the sentence and the second was the action/motion of the agent in the unfolding event. The sentences employed two verb classes–causatives (e.g., *break*) and perception/psychological (e.g., *notice*)–which impose different constraints on the nouns that serve as their grammatical complements. The scenes depicted events in which agents either moved toward a target object (always the referent of the verb-complement noun), away from it, or remained neutral performing a given activity (such as cooking). Scenes and sentences were synchronized such that the verb onset corresponded to the first video frame of the agent motion toward or away from the object. Results show effects of agent motion but weak verb-semantic restrictions: causatives draw more attention to potential referents of their grammatical complements than perception verbs only when the agent moves toward the target object. Crucially, we found no anticipatory verb-driven eye movements toward the target object, contrary to studies using non-naturalistic and static scenes. We propose a model in which linguistic and visual computations in real-world situations occur largely independent of each other during the early moments of perceptual input, but rapidly interact at a central, conceptual system using a common, propositional code. Implications for language use in real world contexts are discussed.

Keywords: situated language processing, visual world paradigm, eye movements, verb meaning, event comprehension, sentence comprehension, language-vision interaction, modularity

Edited by:
Juhani Järvikivi, University of Alberta, Canada

Reviewed by:
Pia Knoeferle, Humboldt University of Berlin, Germany
Laura J. Speed, University of York, United Kingdom

***Correspondence:**
Roberto G. de Almeida
roberto.dealmeida@concordia.ca

†MG passed away before this version of the manuscript was completed

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *07 December 2018*
Accepted: *09 September 2019*
Published: *10 October 2019*

Citation:
de Almeida RG, Di Nardo J, Antal C and von Grünau MW (2019) Understanding Events by Eye and Ear: Agent and Verb Drive Non-anticipatory Eye Movements in Dynamic Scenes. Front. Psychol. 10:2162. doi: 10.3389/fpsyg.2019.02162

INTRODUCTION

How can we talk about what we see? This question, as posed by Macnamara (1978), epitomizes a fundamental problem in human cognition: how we integrate multiple sources of information–different sensory data competing for limited attentional resources–into coherent representations of the surrounding world. The integration between sentences and dynamic real-world scenes, more specifically, depends on a system that can rapidly compute representations of very different kinds. For instance, while the linguistic input system processes phonological, morphological, and syntactic representations, the visual input is dedicated to lines, colors, textures, shapes and other more complex properties of the visual world such as objects and scene layouts. Yet, somehow these types of representations need to "talk" to each other in the process of comprehending what is seen and heard simultaneously. But how do we accomplish such a task within milliseconds of perceiving sounds and lights? How are these seemingly complex and arguably different kinds of representations put together?

We approach these questions by investigating, more specifically, how verbs belonging to different semantic classes, and embedded into sentences, might influence eye-movements to verb-related objects in real, dynamic scenes. In our manipulations, particular objects placed in scenes were always the referents of the complements of the main verbs in the sentences. We employed verbs from two syntactic and semantic classes, one highly constraining regarding the objects it selects in the scenes (causatives such as *crack*, *bend*) and another, non-constraining (perceptual/psychological verbs such as *look*, *inspect*). In addition, we manipulated the nature of depicted events by having agents in the scene moving in different directions–away, toward, or remaining neutral regarding those main objects (see **Figure 1**).

We assumed that the timing of the interaction–that is, the point at which viewers are programming an eye movement to a verb-related object during the speech stream–might indicate the level at which vision and language exchange information. Specifically, if the two systems are encapsulated (or modular, in the sense of Fodor, 1983, 2000; see also Chomsky, 2018), then they should make the products of their linguistic and visual computations available to a common, higher conceptual system. This would result in delayed saccades to objects as a function of verb type and agent-motion type, thus reflecting a late, conceptual interaction. Conversely, if the two systems freely exchange information during their respective early inputs–thus, if they are interactive (McClelland et al., 1986, 2014)–then saccades to verb-related objects should be observed at an early point during verb processing, even in anticipation of the name of the objects attracting saccades, as some studies have suggested (e.g., Altmann and Kamide, 1999; Staub et al., 2012).

These different empirical predictions rely on the widely held assumption that eye movements to particular visual targets are largely under the control of linguistic variables. This assumption comes mostly from studies employing the so-called visual world paradigm (VWP), which involves the concomitant presentation of visual and linguistic stimuli while viewers/listeners have their eyes monitored by an eye tracker (see, e.g., Tanenhaus and Spivey-Knowlton, 1996). A key issue underlying this paradigm is the potential influence of visual context on linguistic processes, as revealed by patterns of eye fixations and scan paths. The use of this paradigm to investigate the architecture of the language-visual interaction raises important questions on the nature of the representations employed in the task of processing sentences and scenes simultaneously. It is of general agreement in cognitive science that for the two systems to influence each other they need to transform their input representations into a common format (Fodor, 1975; Macnamara, 1978; Jackendoff, 1987, 2012). A key question underlying our investigation is whether this common representation format affects early input processing or whether it affects only later stages of processing when both systems have delivered their respective input analyses to a common, higher conceptual system.

In the next section, we first review selected studies that have investigated the interaction between language and vision using the VWP, focusing on the interaction between verb information and objects/referents in scenes or displays. We address methodological issues with these selected studies, which have motivated our empirical investigation employing more naturalistic stimuli. Then, we discuss the proposals for how the two systems might interact. Following this preliminary discussion, we report our eye-tracking experiment involving different verb classes and dynamic scenes. We are particularly interested in how this study can inform us about the architecture that serves the language-vision interaction. To that end, we further develop our proposal for the nature of the representations that afford the interaction between linguistic and visual information in the General Discussion section.

THE VISUAL WORLD PARADIGM AND THE ARCHITECTURE OF THE LANGUAGE-VISION INTERACTION

Visual World Studies

Although the workings of language and vision have been the topic of much research and have figured prominently in integrated models of cognitive processes (e.g., Jackendoff, 1987; Potter, 1999, 2018; Baddeley, 2012) the investigation of how the two systems might influence each other has been somewhat neglected up until recently. While Baddeley's (2012) working memory model, for instance, postulated different buffers for visuo-spatial and phonological processes, it was Jackendoff (1987) who first proposed a model of how visual and linguistic systems might combine their respective representations. But it was the development of the VWP (Cooper, 1974; Tanenhaus et al., 1995) what motivated numerous studies on the interplay between language and vision (see Huettig et al., 2011, for a review). These studies have raised new questions on how visual and linguistic computations interact in the process of understanding what we see and hear simultaneously, with most studies focusing on how information presented in static displays might influence the course of linguistic computations, such as the resolution of temporary lexical, syntactic, or semantic ambiguities. An important advantage of this technique is its ecological validity, for, depending on the variables manipulated, it may closely mimic how linguistic utterances might unfold in visual contexts.

The potential influence of visual contexts on linguistic operations was investigated in Tanenhaus et al.'s (1995) pioneering studies (see also Spivey et al., 2002) involving participants manipulating real objects while following spoken

FIGURE 1 | Sample frames from a dynamic scene accompanied by the sentences *Before making the dessert, the cook will crack the eggs*... (denoting a causative event) or *Before making the dessert, the cook will examine the eggs*... (denoting a perception/psychological event). The three frames represent the onset of three motion conditions of the agent of the event (the cook) with respect to the target object of the sentence (the referent of the *Theme* complement of the verb; the eggs on the kitchen counter): moving toward it, moving away from it, or remaining neutral. See text for discussion, in particular Method (Written informed consent was obtained from the depicted individual for the publication of this image).

commands that contained temporary syntactic ambiguities such as *Put the apple on the towel in the box*. Relevant to our study is the timing of the participants' saccades in relation to the target word (*apple*). In the critical visual conditions, when there was a contrast between one-referent (an apple on a towel) and two-referent contexts (an apple on a napkin and an apple on a towel), the pattern of saccades to and out of the referent region (the apple) were relatively slow. In the one-referent context, subjects looked to the apple about 500 ms after hearing the word *apple*. And in the two-referent context, subjects looked to the correct apple 50% of the time, about 1100 ms after hearing the word *apple*. This study was taken to speak against the view that syntactic structuring–including verb-argument structure–underlying interpretation is an autonomous process, relying primarily on syntactic parsing principles (see, e.g., Fodor, 1983; Frazier, 1988).

This study, however, should be interpreted with caution. First, it had few subjects (six; see also Spivey et al., 2002). Second, subjects were given the opportunity to watch the placement of objects on the table, thus previewing the nature of the incoming displays for "a few seconds" (p. 460), which could have given participants the opportunity to conceive of several possible ways for the task to unfold. More importantly, the timing of saccades reported in the study seems incompatible with the view that subjects are integrating visual context early on during sentence processing. While its possible that delays in the two-object condition may reflect true effects of context on early *parsing* decisions, it is also possible that they reflect late, problem-solving strategies.

Numerous other studies involving this paradigm have manipulated different linguistic variables (words, sentences) and visual materials such as color line drawings (clip art) of scenes (e.g., Altmann and Kamide, 1999; Kamide et al., 2003; Knoeferle et al., 2005) photographs of scenes (Andersson et al., 2011; Staub et al., 2012; Coco et al., 2016), and sets of photographs of people and objects (e.g., Boland, 2005). In the remainder of this section, we restrict our discussion to specific issues on how *verbs* might direct attention to referents of their arguments, but we do so based on a few selected findings that are more directly related to the present study. The relationship between a verb and the referents of its arguments in a scene is of particular concern not only because of its connection with the experiment we report below, but also because it speaks more directly to our main theoretical concern: how representations from vision and language interact such that we can understand and produce utterances referring to what we see.

Altmann and Kamide (1999) employed the VWP to investigate sentence interpretation as a function of different verb selectional restrictions, that is, the types of real-world objects that verbs semantically select (e.g., "something edible" for the verb *eat*). Using ersatz scenes (Henderson and Ferreira, 2004), they found that saccades to the drawing of an object such as a cake were faster when the accompanying sentence was *The boy will eat the cake* than when it was *The boy will move the cake*. Saccades to the picture of the cake occurred on average at about 200 ms (*eat*) and 400 ms (*move*) after the offset of the verb. Eye movements for the constraining condition (*eat*-cake) were said to be *anticipatory*, i.e., they occurred before the noun onset. In their Experiment 1, they found that anticipatory fixations to the object (e.g., cake) occurred in 54% of the trials in the related condition (*eat*), compared to 38% of the trials for the unrelated (*move*) condition. These effects, measured by saccade-to-object onset times and percentage of trials, were obtained when participants were told that they had to judge whether or not objects in the scene were mentioned in the sentence. In the absence of an explicit judgment about sentence/scene match (Experiment 2), the effects were reduced but consistent with the first experiment (32 and 18% in the related and unrelated conditions, respectively). They suggested that what is accessed at the verb "is not structure *per se*, but *interpreted* structure" (p. 259). That is, information about a scene is assimilated by–or interacts with–the ongoing sentence interpretation process such that a *semantically related* argument of the verb has advantage over an unrelated one.

This study was further supported by Staub et al.'s (2012). Employing photographs of scenes, they also found anticipatory eye movements to target objects for semantically constraining verbs (*eat*) but not for control verbs (*move*). There were, however, several differences between the two studies, including method and statistical analyses. For instance, contrary to Altmann and

Kamide, Staub et al.'s (2012) sentences did not contain a 200 ms pause between the verb and the determiner. In Altmann and Kamide's study this could have yielded earlier saccades in the related (*eat*-cake) condition. But it should be noted that Staub et al.'s (2012) results, on the other hand, were obtained with sentences that differed in length: while the relevant constraining segment (*eat the*) took 617 ms, the non-constraining one (*move the*) took 688 ms on average. Assuming that "express saccades" (see Fischer and Ramsperger, 1984) to targets can be obtained at about 100 ms, it is possible that the 71 ms length difference between the two conditions could have given the constraining condition a head start, producing faster saccades to target objects.[1] Indeed, the mean difference in saccade latency between the two conditions–105 ms–is close to their difference in length. Also, percentage of fixations on the target object differed significantly in the two verb conditions, but they were as small as 5% by verb offset and 7.5% by noun offset.

Coco et al. (2016) also offered a quasi-replication of Altmann and Kamide (1999) employing pictures of scenes with sentences containing constraining and unconstraining verbs, such as *The man ate/removed the sandwich*. Additionally, the referent of the internal argument of the verb (*sandwich*) was present or absent in the picture.[2] They did not report latency of saccades, and used part of the scene (e.g., a table) as the region of fixations in lieu of the object. They found a greater proportion of fixations on the "contextual object" (table) with the constraining verb than with the unconstraining one, in a window of time that spans from 100 to 700 ms post verb onset. The difference between the two verb conditions was also significant in the object absent comparison. These results suggest that subjects took the verb information and anticipated a plausible referential location for the upcoming (or unfolding) target object name. It is, however, difficult to contrast this study directly with the others such as Altmann and Kamide (1999) and Staub et al. (2012), for two main reasons: (1) the lack of latency data, and (2) the nature of the data reported, which does not allow us to determine the magnitude of their effects–neither at particular time points after verb onset, nor overall, up to noun onset.[3]

While Coco et al. (2016) controlled for possible word association confounds, it is not clear whether the effects reported in both Altmann and Kamide (1999) and Staub et al. (2012) are due to verb structure (viz., argument or thematic structure) or simply semantic relatedness between verbs and more plausible objects in the context, such as *eat*-cake. This difference is important because if what is at stake is semantic relatedness–a form of priming–it could be argued that the effect does not reflect *influence* of visual context on early linguistic computations, but a late, conceptual effect. Kamide et al. (2003), however, suggest that indeed effects of verb-argument structure are involved in the process of incremental interpretation. In their Experiment 1, employing double object constructions such as *The woman will spread the butter on the bread/The woman will slide the butter to the man*, with scenes depicting the four referents (man, woman, bread, butter) they found that there was a greater proportion of trials in which participants looked at the "appropriate" *Goal* (*bread* in *spread*; *man* in *slide*) than in the "inappropriate" *Goal* (*bread* in *slide*; *man* in *spread*). These effects were not obtained at the verb region (a window of 350 ms post-verb onset) but during the processing of the *Theme*, the noun *butter*. Their verb effects in the *Theme* region, which is a window of 882 ms during the processing of *the butter*, were small (and non-significant, in one of the analyses). This was taken as evidence that listeners anticipate an appropriate *Goal* for a ditransitive verb such as *slide*, but not for a verb such as *spread*. They supported a view of language processing in context that takes into account "all the syntactic, semantic, and real-world constraints that can be applied" (p. 153) at a given segment, attempting to predict the nature of other potential arguments.

Knoeferle et al. (2005) have also investigated the interaction between depicted events and verb-argument processing. In one of their studies (Knoeferle et al., 2005), they presented three characters (e.g., a princess, a pirate, and a fencer) as an ersatz scene while participants heard sentences referring to their roles (e.g., *The princess is apparently washing the pirate* [Subject-Verb-Object]/*The princess is apparently painted by the fencer* [OVS]). In their Experiment 1, they found greater inspection (proportion of looks) of the object (e.g., the pirate) upon hearing the verb. The effect was obtained at about 2000 ms from the onset of the sentence and before the grammatical object was uttered. Notice, however, that the scenes had only three characters: the one in the center was the princess, which could be understood as subject or object, thus there was a 50% chance that one of the two remaining characters would be looked at once the princess was inspected. Proportions of looks to the appropriate character, however, were below chance during the verb, with about 40% of the looks into the princess up until the adverb onset. This study was taken to support the idea that "non-linguistic information–such as contrast, actions, or events–that establishes relevant relations between entities, can affect how linguistic input is interpreted" (p. 122).

In summary, the studies briefly reviewed in this section claim to provide strong support for the view that information about the visual context aids linguistic processes of syntactic structuring (argument/thematic structure) and semantic interpretation. The language comprehension system is said to be incremental, at each moment considering *all* available sources of information–and in particular, verb-related information such as the conceptual nature of arguments. And it is because language use normally occurs in visual contexts that those studies supposedly carry a high

[1] It should be noted that express saccades are obtained under very different conditions (e.g., after the onset of a dot in peripheral vision) and, to our knowledge, have not been investigated with more complex but static naturalistic stimuli such as those of Staub et al. (2012). Our suggestion, however, is that saccades to targets *could* occur at a latency that is close to the difference in length between the conditions in Staub et al.'s (2012) study.

[2] We only briefly discuss–and hereafter refer to–Coco et al.'s (2016) Experiment 1 because their Experiment 2 employed a "blank screen" paradigm with a scene preview of 5000 ms, thus constituting a memory experiment.

[3] Regarding (1), we deem this necessary to understand the reflexive (or lack thereof) saccades to target, which might signal the strength of the context by verb type interaction (see below, in Results, our analyses of this measure). Regarding (2) it is quite possible to obtain a difference in magnitude of proportion of fixations to the target with a very small number of fixations (see Kamide et al., 2003, for one example). Although this is not necessarily the case in Coco et al.'s (2016) study, to claim anticipatory effects from such data might lead to a generalization that stands on the exception rather than on what might be the rule.

degree of ecological validity, bearing on the nature of cognitive processing architecture: the weight of the evidence seems to favor a highly interactive and probabilistic rather than a modular, rule-based view of language comprehension in visual contexts.[4] However, as we pointed out, there are numerous methodological issues with these visual world studies, casting doubt on their generalizability.

Problems With the Visual World

One of the main problems with the visual world studies reviewed above is the timing and nature of saccades reported. For the most part, saccades to relevant objects are relatively late, when other linguistic constituents are being processed (as in Tanenhaus et al., 1995; Spivey et al., 2002); or they occur after an artificially introduced break in the sentence (Altmann and Kamide, 1999); or they are triggered by linguistic segments of unequal length (Staub et al., 2012). Moreover, the proportion of fixations that are usually reported as evidence of anticipatory eye movements to targets is relatively small, often below chance. Consider, for instance, Kamide et al.'s (2003) Experiment 2. They report anticipatory effects in a region spanning 637 ms after verb onset with looks to relevant targets occurring in only 10% and 7% of the trials in the experimental and control conditions, respectively (a statistically significant effect). What is perhaps most surprising in this study is that looks to other regions of the scene occur in 55% of the trials with agents receiving 35% of all fixations, an amount greater than any of the target objects they were contrasting. Thus, although verb-scene semantic effects were found, most eye movements do not appear to be locked into the initial process of interpreting arguments and their referents in the visual world.

A second methodological problem with these studies is the nature of the visual context they use. Given that they do not involve realistic dynamic scenes, they can only generalize to language use in static contexts, limiting the strength or their challenge against modular systems. One possible interpretation of their results is that the lack of agents and motion in depicted events frees attentional and gaze mechanisms to be controlled by linguistic processes of interpretation, thus yielding effects of anticipation. Notice that this does not rule out that early interference of visual context on linguistic processes might be exerted–but they might occur under relatively artificial conditions, thus weakening the claims that vision and language are interactive tout court. Moreover, given the timing of saccades and the small proportion of early *fixations* (which occur after targets are encoded and saccade programming occurs) the effects obtained by visual world studies are compatible with a view that takes language and vision to be initially encapsulated, but interacting at higher processing levels (see below). Knoeferle et al.'s (2005) visual contexts, for instance, require participants to *infer* that a princess would be holding a bucket with the purpose of washing a pirate; or that a princess would be holding an artist's brush and palette to be painted by a fencer. Not only do the scenes look unnatural, the level of detail required to make the appropriate inferences about the characters' roles may account for the high percentage of fixations (about 40%) to the princess, the central character, during verb processing and up to adverb onset.

A third reason for questioning the claims stemming from the visual world literature is the nature of the representations and the interaction mechanisms proposed. The main assumption is that visual and linguistic processes interact early on, possibly at perceptual levels of analyses. In fact, most studies supporting interactionism have lined up with a cognitive architecture that postulates no clear distinction between levels of analysis or processing components (e.g., Altmann and Kamide, 2007; Mayberry et al., 2009; Smith et al., 2017). Smith et al. (2017), for instance, take the constraints exerted between semantic and visual processes over linguistic/phonological processes to be a function of co-activation of nodes in a connectionist network. But, this only begs the question about what nodes stand for and how links are obtained in the "integrative layer" (viz., how a "phonological" node talks to a "visual" node). Besides these nodes being arbitrarily determined in terms of both the *content* they stand for and the nature of the links that obtain between them at different layers, a perennial problem with connectionist models of the semantic system is that they cannot account for the *compositionality* and *productivity* characteristic of linguistic and conceptual systems (see Fodor and Pylyshyn, 1988). Mayberry et al.'s (2009) CIAnet is also a connectionist model based on Knoeferle and Crocker's (2006) coordinated interplay account (CIA, for short). It incorporates attentional mechanisms that are responsible for activating "the event in the scene most relevant to the utterance" and it does so by "learning to bind the events' constituents" (Mayberry et al., 2009, p. 462). While CIA (and CIAnet) is explicit about some of the representations computed (the likes of *Agent*, *Action*, *Patient*), it does not make explicit how vision extracts that information from the scene, other than assuming that visual inspection produces event interpretations and predictions. Contrary to these models, the architecture we propose below is committed to a common representation format for the interaction between language and vision, relying on the independent and parallel computation of both systems at the earliest stages of processing during dynamic scene and sentence comprehension.

The Nature of Representations and the Architecture of the Language-Vision Interaction

In the present study, we manipulated realistic dynamic scenes and sentences aiming to understand the interaction between linguistic and visual representations in real-world situations. Consider, for instance, witnessing one of the events depicted in **Figure 1** while listening to a sentence such as *The cook will crack the eggs that are in the bowl*. Upon hearing the causative verb *to crack*, there are only a few objects in the scene that

[4]It is important to note that the notion of interactive system usually proposed in the reviewed visual world studies assumes that modular systems allow for *no semantic* information guiding lower linguistic processes such as sentential parsing and word recognition. However, a recent proposal for incorporating semantics into the linguistic module (de Almeida and Lepore, 2018) would in part account for semantic effects in linguistic computations–viz., by regulating local discourse elements and possibly visually computed referents–to aid in the construction of a semantic representation, before other forms of knowledge (e.g., background/world knowledge) influence linguistic decisions.

might be relevant for understanding the unfolding event–the *eggs* among them. Consider now the same sentence with a perceptual/psychological verb such as *to examine*. While *crack* restricts the potential referents in the scene, *examine* allows for a wider range of objects, as possibly anything in the scene can be examined. Moreover, while the causative class denotes a change of state in an object in the scene (the referent of the *Theme*), the psychological class might denote a change of state in the agent (the *Experiencer* of the event). Therefore, these verb types represent a clear contrast with regards to the relation between an agent (typically, the sentence's grammatical subject) and a real-world object. As we have seen above, several visual world studies using ersatz or static pictures of scenes have suggested that verbs lead to *anticipatory* eye movements to the referents (i.e., the objects) of their noun complements when a given relation between a verb and potential referent is established. Although there are numerous methodological differences between these studies, as well as different interpretations for their results, the effect has not been generalized to realistic, dynamic scenes, which arguably better represent "how we talk about what we see."

One of the problems with static scenes is that they are predictable and thus might not tax the attentional system, allowing the eyes to move freely, promoting eye movement behavior that is more likely to be in consonant with the linguistic utterance (but see Andersson et al., 2011). Dynamic scenes, however, while often predictable due to physical constraints (e.g., inertia and gravity), also have a high degree of unpredictability in particular when human agents are involved. As shown in **Figure 1**, there are numerous possibilities with regards to what the agent might do while remaining true to either the *...crack...* or *...examine...* versions of the sentence. We can assume that at least three actions might take place: the agent may move toward the object, she may move in the opposite direction, or she may remain neutral (i.e., continue mixing the dough). Would different agent behaviors affect what the viewer attends to? And, more specifically, would attention be drawn to target objects (*eggs*) independent of and in anticipation to agent action in the case of the more restrictive causative verb? If, as studies have suggested, attention is driven to objects *automatically* and in *anticipation* of the noun being heard, we should expect that agent action should not affect saccades to their selected objects; along the same line, we should expect a verb effect to be obtained, with the more constraining verb (*crack*) always leading to faster saccades to the referent of its complement noun (*eggs*).

While these questions about the role of verb restrictions and agent action in dynamic scenes are important for understanding how linguistic and visual processes might influence each other, equally important are the processes that allow for the combination of representations from both input streams. In Jackendoff's (1987, 2012) model, linguistic and visual inputs run parallel, independent processes analyzing their respective stimuli. Their outputs reach a central, conceptual structure system, after they are translated via interface modules and coded into a propositional form, compatible with both, the representation of the spatial structure on the visual side and the representation of the structural properties of the sentence, on the linguistic side.[5] This conceptual system, more importantly, operates on a symbolic, amodal code that is common to the *products* of both input systems, language and vision.

Our hypotheses stem from a parallel modular architecture that is closely aligned with Jackendoff's (2012) model. As **Figure 2** shows, we assume two main autonomous input systems that feed a central conceptual system, which dynamically updates accessed representations in conceptual short-term buffer (CSTB). CSTB actively combines concepts into propositional structures computed from both linguistic utterances and dynamic scenes. While predicates and arguments constitute the basic building blocks of linguistic-semantic representations, they have also been proposed to constitute the fundamental representations of visual processes (Pylyshyn, 2007). These predicates are primarily descriptive of object and scene properties (the "arguments") such as spatial relations and scene dynamics (e.g., trajectory of objects

[5]See Jackendoff (2012) for details on how conceptual structure differs from other cognitive systems as those responsible for computing visual-spatial structure. See also Jackendoff (1990) on how concepts–the elements of conceptual structure–might be themselves structurally complex representations. In the present discussion, we gloss over the technical details of this proposal, in particular with regards to whether or not concepts are complex (see de Almeida and Manouilidou, 2015, and de Almeida and Antal, in press, for discussions on how concepts lexicalized by verbs might be represented).

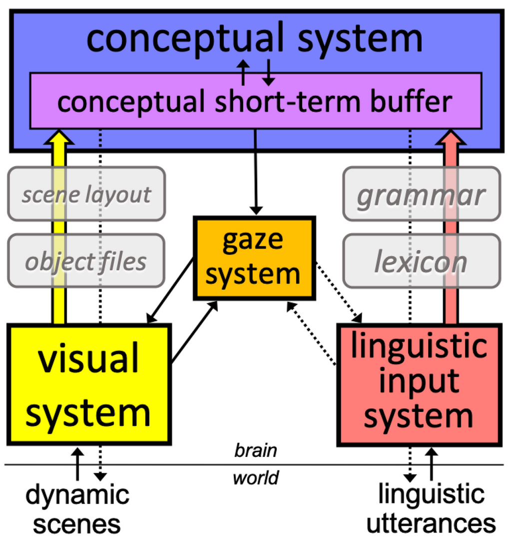


FIGURE 2 | A schematic representation of the proposed modular processing architecture bearing on the interaction between language and vision. The solid arrows represent the mandatory flow of information from input systems to the conceptual system and from this to the gaze system. The dotted arrows represent hypotheses about top-down attentional mechanisms involved in (a) locking into referents in the world (e.g., tracking objects, updating representations), (b) attending to the spoken source, (c) tracking low-level properties of linguistic analyses (e.g., computing phrase structure), and (d) reading mechanisms (e.g., word fixation and next-word saccade programming). The conceptual short-term buffer integrates predicate-argument relations computed from both systems, relying on the input systems' knowledge of objects and scene layouts, on the visual stream, and the lexicon and the grammar, on the linguistic stream. See section "General Discussion" and **Supplementary Figure S2** for a more detailed presentation of the workings of the model.

and agents). The gaze system programs and executes saccades based primarily on input from the visual system (exogenous), and conceptual-structure representations (endogenous), including schemas such as those deployed in the control of voluntary activities (see, e.g., Land, 2009). The gaze system also responds to linguistic input, in particular lexical properties, which determine patterns of saccades and fixations in reading (Reichle et al., 2003). Crucial to the interaction between the two systems is the allocation of attention to scenes and to the linguistic input. Visual attention locks into multiple objects in the visual field, which are the primary sites for potential fixations during scene and sentence processing.

While details of this model's operations are beyond the scope of the present introduction (see section "General Discussion" and **Supplementary Figure S2**, for further description), it is important to highlight its predictions for the simultaneous processing of visual and linguistic representations.

First, we predicted that the more restrictive type of verb, causative, would lead to faster saccades to the object, compared to the less restrictive perception verb. We assumed that if there are anticipatory saccades to the target object, the nature of the object that matches the semantic restrictions of the verb has to be already encoded or activated. This would imply that viewers/listeners would have established early on a connection between verb meaning and object meaning. This connection would also reflect in anticipatory saccades to target objects across scene types, but with greater effects in the toward condition, compared to neutral and away conditions. By contrast, a lack of anticipatory eye movements could be seen as the eye-gaze system's reliance primarily on visual processes of scene comprehension, over and above the potential effects of verb-semantic restrictions. In other words, one should expect saccades to the target object to reflect a late assessment of *both* visual context and verb-semantic restrictions, such that only upon hearing the referent of the internal argument *and* evaluating the scene semantics that saccades to target objects would occur. This prediction would rule out verb-driven anticipatory effects, but not motion effects, with toward condition yielding faster but non-anticipatory effects to the target, independent of verb type. We should note that, as pointed out by a reviewer, it is possible to conceive of a late interaction to be driven by factors such as the cognitive load brought about by the complexity of the scene (see, e.g., Andersson et al., 2011) in lieu of architectural restrictions. We contend, however, that related studies using similar naturalistic scenes as ours (e.g., Staub et al., 2012), though without motion, did report anticipatory eye movements. If scene motion is a factor we should expect an interaction between verb type and agent motion, signaling verb-driven effects on eye movements to related objects with a causative-verb advantage. Moreover, our toward condition should *promote* anticipatory eye movements to the referents of causative verb complements. This is so because, contrary to other studies with static scenes, participants have two sources of information to drive their potentially anticipatory looks: verb restrictions and agent motion. Ultimately, the present study addresses a methodologically important point which is the reliability of the VWP to make claims about the architecture of the language-vision interaction insofar as realistic motion events are concerned.

METHODS

Participants

Thirty-eight Concordia University students (32 females) participated in the eye-tracking experiment, none of them participated in the norming studies developed for the preparation of materials. Participants were all native speakers of English and had normal or corrected-to-normal (with contact lenses) vision. Their age ranged between 19 and 38 years of age ($M = 22$; $SD = 3$). They all gave written informed consent and participated for course credit as part of the Concordia Psychology Participant Pool. The experiment was approved by the Concordia University Human Research Ethics Committee.

Materials and Design

Verbs

We initially selected 18 lexical causatives (e.g., *crack*, *bounce*), and 18 verbs that were either perception (e.g., *see*, *hear*) or psychological verbs (e.g., *examine*, *notice*), following Levin's (1993) classifications. Only 17 of each kind were used in the eye-tracking experiment due to a movie recording error which led us to eliminate one trial before we conducted the experiment but after all materials had been prepared. Norms for verb complements were obtained from a study with 58 Concordia University participants, all native speakers of English, who did not take part in the main experiment. These participants were required to fill-in simple sentence frames such as *The man bounced the* __________ and *The* _________*bounced*. The complements used in the sentences and movies were chosen based on three criteria. (1) They were the nouns most frequently given in the frames. (2) They were not the strongest associates of the verb. This association was determined by 10 other participants who were required to provide the first word that came to mind for each verb. Thus, if for a frame such as *open-*_____ the most frequent associate provided was *door*, we eliminated *door* as a complement for *open*. And, (3) the referent objects constituted mid-size objects, which could be filmed in mid-ground. Therefore, we eliminated cases such as *wall* for *crack* (the chosen was *eggs*). The perception and psychological verbs were also determined based on Levin's (1993) classification of verbs that take an *Experiencer* as external, subject argument, as opposed to an *Agent*. This criterion applied to at least one sense of those psychological verbs that can have multiple senses. These verbs were matched with the causatives based on frequency (Kucera and Francis, 1967; Coltheart, 1981) but also on the plausibility of the events to be filmed (to allow for pairs such as *. . .crack/examine the eggs*), as judged by the experimenters and two research assistants.

Sentences

Seventeen sentence pairs were created, with each member of a pair differing only with respect to the main verb, which belonged either to the causative or the perception/psychological class (e.g., *Before preparing the dessert, the cook will crack/examine the eggs that are in the bowl*). All sentences had an initial patch clause, which was always of an adverbial type (e.g., *Before preparing the dessert…*, *After playing with the toys…*), followed by a main clause. All main clauses were of the form NP1 (Noun Phrase1)-*will*-Verb-NP2-RC (Relative Clause). The NP1 always referred to the agent in generic form (*the cook*, *the boy*, *the man*, etc.); the NP2 referred to the target object in the scene (*eggs*, *ball*, etc.); and the RC always made reference to the target object (e.g., *…that are in the bowl, …that is on the bench*). It should be noted that although our sentences contain a RC with a prepositional phrase that could signal a potential disambiguation between two competing referents, this comes after the crucial complement noun, thus after the point of interest regarding verb-driven saccades to target objects. Moreover, as can be seen in our materials, in most of our scenes there is only one referent for the verb complement noun.[6] Sentences were recorded by a female research assistant speaking at a normal pace.

Scenes

Each movie consisted of single shots of about 10 s of indoor or outdoor naturalistic scenes. There was no camera movement or zoom, and the only source of motion within the movies was that of the agent performing a given action. For films produced at a large furniture store, the store displays were arranged to resemble common household areas (e.g., kitchen counters were filled with utensils, bookshelves were filled with books and objects). For all the films produced at houses, parks, and streets, the only scene alterations were the placement of target objects (e.g., a kite on a park bench). Agents and target objects were on the same plane (always mid-ground) in opposition to each other (e.g., if the agent was on the left, the target object was on or near the right edge of the visible image). Each of the 17 unique scenes (e.g., someone cooking in a kitchen) was filmed with three different endings: after an initial similar segment of about 7 s, agents moved (or reached) either toward a particular target object, away from it, or remained neutral, that is, continued doing what they were doing in the initial segment (see **Figure 1**). There was thus a total of 51 unique movies (17 scenes × 3 endings). Each movie was then synchronized with two sentences (causative and perception/psychological), yielding 102 film/sentence combinations. The 102 stimuli were distributed in six lists of materials, each one containing 17 trials (film/sentence combinations), with two or three of each verb-type/motion-type combination. We did not use filler items noting that Staub et al. (2012) have obtained no statistically significant differences between lists of materials when they were ran with different filler types (e.g., standard, unpredictable) and no fillers. Film resolution was set at 720 × 480 pixels in NTSC format (29.97 frames per second). The digital movies were produced and edited by a Concordia film student using Final Cut Pro (Apple, Inc.). Scene norms included object saliency and scene semantics (predictability of events). These are described in **Supplementary Material**. Please also see **Supplementary Material** for sample video (written informed consent was obtained from the depicted individual shown in the video for its publication).

Sentence-Film Synchronization

Sentences and films were synchronized such that verb onsets in the sentences corresponded to agent action onsets in the movies. The agent action onset was determined by inspecting, frame-by-frame, the point in which the agent started moving toward or away from the target object. For each movie, there were up to seven frames in which the agent motion direction could be said to have started (e.g., beginning of rotation of the torso or limbs toward or away from target object). The investigators selected one of these frames as the onset of the action. Since each frame corresponds to 33.37 ms, the onset of the agent motion–that is, the moment the agent turns unambiguously toward or away from target object–was determined within a window of 234 ms (7 frames), with the corresponding time in the neutral motion condition. We call this the “disambiguating point” (see also **Supplementary Figure S1**). Acoustic onsets of the verbs in the sentences were determined by amplifying the acoustic waveforms in the digital sound files and identifying the lowest frequency marking the boundary between words, or by splitting transitional phonemes when the lowest frequency was not obtained.

Apparatus and Procedure

Visual and auditory stimuli were presented via PsyScope software (Cohen et al., 1993) on an Apple computer placed 41 cm away from participants. Participants wore clip-on headphones. Their eye movements were recorded using the EyeLink-I head-mounted eye-tracker at a sampling rate of 250 Hz from the left eye only (viewing was binocular). Head movements were minimized with the use of a chinrest. **Figure 3** shows schematically how stimuli were presented in each trial, from the fixation cross to the verb and agent activity onset. Participants were instructed to watch the movies and to listen to the accompanying sentences while paying close attention to both as they would later be tested with a recall task. They were asked to press the space bar on the computer keyboard to begin each trial. Each trial lasted approximately 12 s and the full experimental session lasted about 30 min, including eye-tracker calibration time.

RESULTS

Data Integrity and Analysis Methods

The post-experiment recall test consisted of a booklet with 12 sentences and 12 still frames of the movies, half of which had been presented in the experiment. Participants scored between 75 and 100% (M = 87.6%, SD = 7.3%) in the cued recall task, indicating that they attended to both movies and sentences; therefore, data from all participants were kept in the analyses. Of the total trials, 4.5% were missed due to corrupted data. These trials were distributed evenly across the various experimental

[6]A reviewer also pointed out that clauses with *Before…* have been found to affect working memory because, contrary to *After…* listeners/readers are required to alter the temporal order of events described in upcoming clauses, as suggested by an ERP study conducted by Münte et al. (1998). The assumption is that *Before* sentences could contribute to delaying saccades to target object. However, of our 17 sentences, only four began with *Before*, and three with *After*. See **Supplementary Material** for all sentences and scenes.

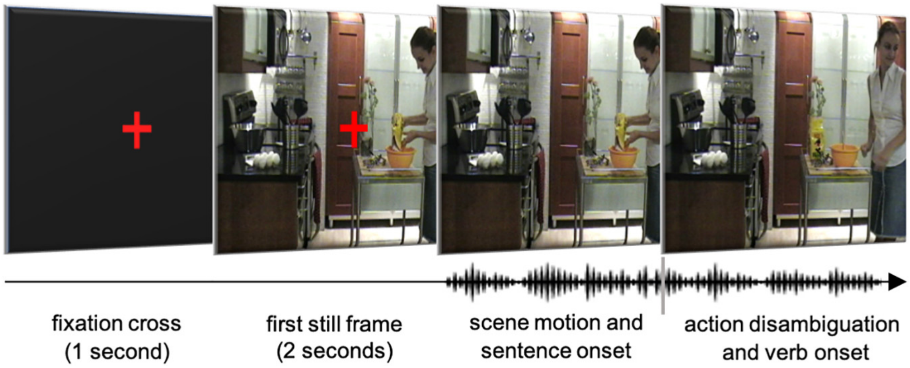


FIGURE 3 | Schematic depiction of the procedure. At the beginning of each trial, participants were presented with a red cross over a black background, centered in the middle of the screen. After 1 s, the black background was replaced by the first still frame of the movie for 2 s, with the fixation cross still in place. After that point in time, the cross disappeared and the movie was set in motion. The accompanying sentences began within a few seconds after the movies began, but the actual onset time of each sentence varied, depending on the synchronization between the acoustic onset of the main verb and the frame corresponding to the beginning of the action performed by the agent in the movie–which was different for each movie triplet (Written informed consent was obtained from the depicted individual for the publication of this image).

conditions. In 91 trials (14.3%) participants did not look at the target object. A 2 (verb type) × 3 (motion type) repeated-measures ANOVA conducted on these trials showed no effect of verb type ($F(1,74) = 1.81$, $p = 0.19$) or of motion type ($F(2,74) < 1$, $p = 0.88$), nor was there a significant interaction effect between the two main variables ($F(2,74) < 1$, $p = 0.52$). This indicates that the apparent motion of the agents in the scenes and the verb class did not affect whether or not participants fixated on the target object after verb-onset; in other words, the number of trials with no post-verb fixations was evenly distributed across the six conditions. The third source of missing data derived from trials in which participants happened to be fixating the target object at verb onset. This occurred in 41 (6.4%) of the trials. These trials had to be excluded from any analyses examining the effect of verb type on subsequent eye movement behavior because it was not possible to determine whether or not participants continued to fixate the object as a function of the verb that they heard.[7] Again, a 2 × 3 repeated-measures ANOVA was conducted to examine the effect of verb type and motion type on the proportion of these trials (by subjects). The results indicated that the verb type failed to reach significance, $F(1,74) = 3.84$, $p = 0.06$, although there was a tendency for a larger proportion of trials to be of the causative type across all motion conditions. But there was no main effect of motion type, $F(2,74) < 1$, $p = 0.64$, nor a significant interaction effect, $F(2,74) < 1$, $p = 0.98$, as expected. These results suggest that the agent's apparent direction of motion at verb onset did not affect whether participants were fixating the target object at the time the verb was spoken.

Analyses of the remaining data took into account three basic measurements of eye movement behavior. (a) Saccade onset time (SOT) which was the time taken by the viewer to launch an eye-movement (saccade) to the target object (always the object referent of the grammatical complement of the main verb of the sentence) after the disambiguating point. This measurement took into account all saccades and fixations made between verb onset and the final direct saccade to the target object but without counting the target fixation time; for the causative sentences, this corresponded to a window of 666 ms and, for the perception sentences, 768 ms. We took this analysis to represent a more on-line measurement of the reflexive behavior triggered by attentional grabbers in the visual scene (directed or not by linguistic cues), in comparison with often reported proportion of trials in which a fixation to a target occurred. (b) Number of saccades produced after the disambiguating point until the viewer fixated on the target object; this was the total number of saccades from verb onset including and up to the first saccade to the target object; and (c) cumulative saccades to the target object during the period following the verb-onset up to the acoustic offset of the spoken noun (the direct object of the main verb).

We relied on two data-analytic methods. First, we report repeated-measures ANOVAs taking into account participants (*F1*) or items (*F2*) as random variables, together with planned pairwise comparisons between verb types across different levels of the motion condition, also considering either participants (*t1*) or items (*t2*) as random variables. For these analyses, data from seven participants who had missing values (means) in one or more conditions were removed from all analyses. Following the removal of the seven participants, we then screened the data for potential outliers. All SOTs below 200 ms were subsequently removed, which resulted in the removal of 23 data points (4% of the total dataset). Furthermore, the Shapiro–Wilk's test revealed that the assumption of normality had been violated for four conditions (i.e., causative-away, causative-toward, perception-away, perception-neutral) in the

[7]Notice that some of the visual world studies discussed above (e.g., Altmann and Kamide, 1999; Staub et al., 2012) included in their analyses trials in which participants were already looking at the target object at verb onset. They took the next fixation as "the first valid fixation on the target, as this is the first fixation that could, in principle, be guided by the selectional restrictions of the verb" (Staub et al., 2012, p. 932). We notice that this occurred on average in 21% of the trials in Staub et al.'s (2012) study, and in 10% of the trials in Altmann and Kamide's study, compared to 6.4% in the present study.

SOT data only. Thus, for all analyses we conducted the ANOVAs on the logged transformed data. Second, we report linear mixed effects models (LME) for all three eye-movement datasets, also with subjects and items as random effects. Although the latter analysis method has become standard in psycholinguistic experiments, we deemed the first analysis type particularly important for it would allow us to contrast our effects with those of previous studies employing the VWP with scenes and different verb types (in particular, Altmann and Kamide, 1999, and Staub et al., 2012).

For the LME analyses (Baayen et al., 2008) we employed the lme4 package (Bates et al., 2013) for the R statistical programming environment (R Development Core Team, 2012). Given that LME accounts for participant idiosyncrasies, we entered all of the participants' raw data into the analyses. As such, the seven participants that were originally excluded in the ANOVA's analyses, due to missing values in one or more conditions in the SOT data, were included, resulting in a total of 38 participants.

For all LME analyses, our models were fitted using a backward step-wise elimination procedure, whereby the predictor variables that did not significantly improve the model as indicated by likelihood ratio testing were subsequently removed. Furthermore, all models included only random intercepts for participants and items, as justified by the likelihood tests, given that the simple model could not be rejected in favor of a more complex model. The most complex model, which included random slopes for the fixed effects and their interaction, did not converge for any of our analyses, and thus could not be evaluated. We then derived *p*-values for each predictor variable by comparing the fitted model to a minimally contrasting null model that excluded the relevant term. Planned comparisons were conducted using the emmeans package (Lenth et al., 2018).

Saccade Onset Time

Figure 4 depicts the mean SOT for each verb and motion condition. There was a main effect of motion type ($F1(2,60) = 11.27$, $p < 0.0001$, $\eta_p^2 = 27.3$, observed power = 0.99; $F2(2,32) = 9.75$, $p < 0.0001$, $\eta_p^2 = 37.9$, observed power = 0.972), and a marginally significant effect of verb type in the participants analysis ($F1(1,30) = 4.16$, $p = 0.05$, $\eta_p^2 = 12.2$, observed power = 0.506), and in the items analysis ($F2(1, 16) = 4.14$, $p = 0.059$, $\eta_p^2 = 20.5$, observed power = 0.481). There was no significant interaction between verb type and motion type ($F1(2,60) = 0.31$, $p = 0.73$, $\eta_p^2 = 0.10$, observed power = 0.10, $F2(2,32) = 0.46$, $p = 0.63$, $\eta_p^2 = 2.8$, observed power = 0.12). Planned comparisons between verb types across different motion conditions revealed that causatives yielded faster SOTs than perception/psychological constructions only in the toward condition for items analysis ($t2(16) = 1.80$, $p = 0.047$), but not for subjects. No other comparisons between verb types across different levels of the motion variable reached significance. The magnitude of SOTs in the toward condition was similar to those obtained in studies that employed a similar methodology but with static pictures/drawings of scenes, such as Kamide et al. (2003), although larger than those of Altmann and Kamide (1999). The lack of a verb effect in the away and neutral conditions is surprising. Allied to the motion condition main effect, the lack of verb effect in the away and neutral conditions suggests that eye-movements are controlled primarily by the agent motion, with saccades to target objects in the away and neutral conditions showing *insensitivity* to verb-thematic properties. In addition, there were no *anticipatory* effects, since saccades to the target object occurred 140 ms *after* the noun offset in the fastest causative-toward condition (in the slowest case, the perception-neutral condition, the SOT to the target occurred 678 ms after the noun offset). Thus, even in the fastest condition, with constraints posed both by the action of the agent (moving toward a given object as opposed to moving away from the scene or remaining neutral) and by the highly constraining causative verb, programming of the saccade may have occurred within the noun object. This interpretation takes into account the best estimate of a 200 ms attentional shift preceding the saccade, as made in other static-scene visual-world studies.

For the LME analyses, SOT to target was entered as the dependent variable, motion type and verb type were entered as fixed effects, and participants and items as random predictors. The SOT model was compared to a null model consisting of only random predictors and was found to provide a better fit to the data, $\chi^2(5) = 84.83$, $p < 0.001$. We also found a main effect of motion type, but no main effect of verb type or interaction (see **Table 1**).

Multiple comparisons revealed that there was a significant difference in SOTs between the toward and away motion type conditions, $p = 0.001$, and the toward and neutral conditions, $p < 0.001$, but no difference between the away and neutral conditions, $p = 0.264$, suggesting that participants are faster to look at the target object when the agent was in the toward condition. Furthermore, planned comparisons revealed no statistically significant differences between causative and perception verb types on all three levels of motion type (all $p = 0.79$), in contrast with the pairwise analyses we performed employing *t1* and *t2* separately. We thus suggest that there is a weak, though non-anticipatory effect of verb type (numerically faster causatives) in the toward condition.

Number of Saccades to Reach the Target Object

For number of saccades (**Figure 5**), a similar pattern of results was obtained: main effect of motion type ($F1(2, 60) = 9.07$, $p < 0.0001$, $\eta_p^2 = 23.2$, observed power = 0.969; $F2(2, 32) = 5.88$, $p = 0.007$, $\eta_p^2 = 26.9$, observed power = 0.842) but no main effect of verb type ($F1(1, 30) = 0.37$, $p = 0.55$, $\eta_p^2 = 1.2$, observed power = 0.091; $F2(1, 16) = 2.12$, $p = 0.17$, $\eta_p^2 = 11.7$, observed power = 0.278). There was a significant interaction between verb type and motion type ($F1(2,60) = 7.81$, $p = 0.001$, $\eta_p^2 = 20.7$, observed power = 0.942, $F2(2,32) = 0.247$, $p = 0.78$, $\eta_p^2 = 1.5$, observed power = 0.09). The number of saccades within the scene until the subject reached the target object for the first time, counting from verb onset, was relatively small–from 2.69 in the causative-toward condition to 4.2 in the perception-neutral condition. In the present study, the number of saccades also suggests that fixations were fast (in the magnitude of 298 ms each for the faster causative-toward condition). Planned comparisons revealed that

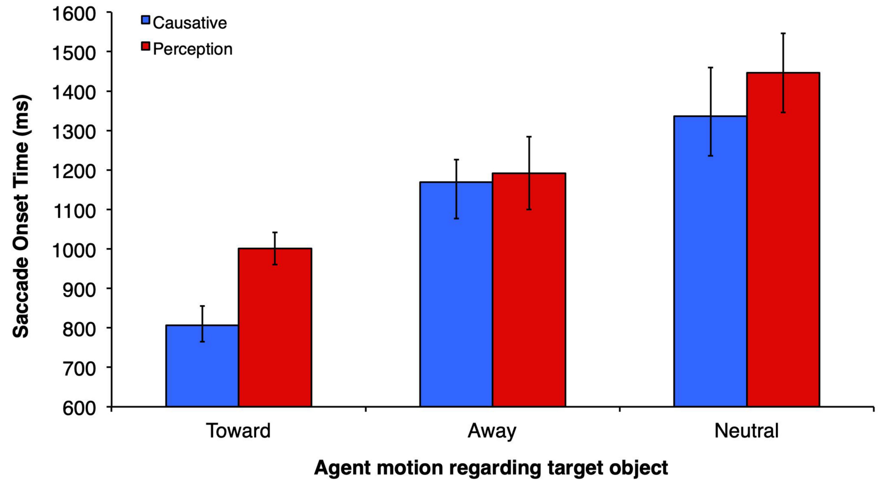


FIGURE 4 | Saccade onset time (SOT) representing the mean time taken by viewers to launch a saccade to the target object after the onset of the verb.

TABLE 1 | Linear regression of Saccade Onset Time.

Predictor	β	SE β	*t*-value	95% CI of β	Null comparison
Constant	1402.64	125.90	11.14	[934.08, 1923.87]	
Motion type	−136.67	35.71	−3.83	[−368.85, 69.94]	$\chi^2(2) = 29.73$, $p < 0.001$
Verb type	79.86	57.92	1.38	[−249.01, 373.07]	$\chi^2(1) = 1.67$, $p = 0.197$
Motion type × Verb type	8.605	71.23	0.12	[−131.01, 148.22]	$\chi^2(2) = 0.08$, $p = 0.961$

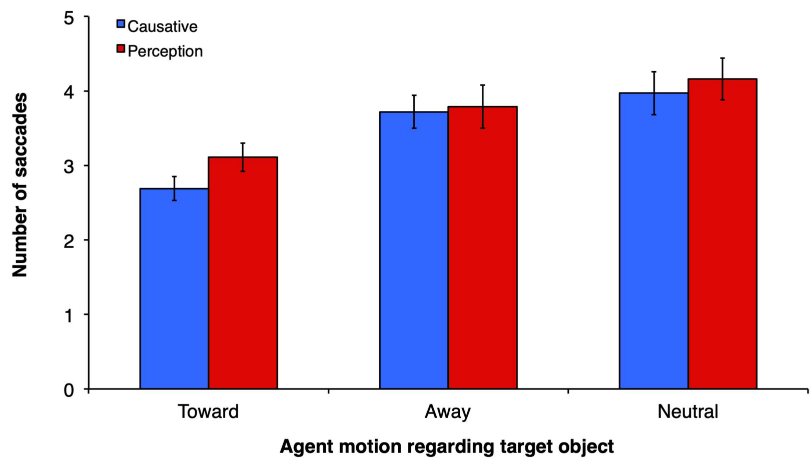


FIGURE 5 | Mean number of saccades from the onset of the verb until the time participants fixate on the target object (grammatical complement of verb).

in the toward motion condition the number of saccades to the target object was smaller when the verbs were causatives than when they were perception/psychological ($t1(30) = 5.23$, $p < 0.0001$; $t2(16) = 2.62$, $p = 0.011$). No other comparisons between verb types across motion conditions were significant.

For the LME analyses we used number of saccades from verb onset until object reached the target as the dependent variable. Motion type and verb type were entered as fixed effects, and participants and items as random predictors. This model was compared to a null model consisting of only random predictors and was found to provide a better fit to the data, $\chi^2(5) = 35.54$, $p < 0.001$. Similar to our SOT LME analyses, we found a main effect of motion type, but no main effect of verb type or interaction (see **Table 2**).

Multiple comparisons indicated that there was a significant difference in number of saccades between the toward and away motion type conditions, $p < 0.001$, the toward and neutral conditions, $p < 0.001$, but no statistically significant difference between the away and neutral conditions, $p = 0.35$, indicating that participants performed less saccades before reaching the

TABLE 2 | Linear regression for total number of saccades before reaching target object.

Predictor	β	SE β	*t*-value	95% CI of β	Null comparison
Constant	2.73	0.53	5.19	[3.14, 6.18]	
Motion type	0.58	0.10	6.07	[−1.34, −0.11]	$\chi^2(2) = 34.89, p < 0.001$
Verb type	−0.22	0.29	−0.77	[−1.10, 0.87]	$\chi^2(1) = 0.68, p = 0.411$
Motion type × Verb type	0.12	0.19	0.62	[−0.21, 0.57]	$\chi^2(1) = 0.39, p = 0.534$

target object when the agent moved toward that object. Planned comparisons revealed no statistically significant differences between the causative and perception verb types on all three levels of motion types (all $p = 0.96$). Numerically, however, the number of saccades to reach the object in the causative condition was smaller (2.69) than in the perception condition (3.11). The magnitude of these values is similar to those obtained by Staub et al. (2012), in their restricting (2.77) and control (3.23) conditions.

Cumulative Saccades to the Target Object

Our time-course analyses relied on cumulative saccades to the target object, rather than the usually reported proportion of fixations per trial. As we have seen in the review of visual world studies, proportion of saccades divided by trials and time slots tend to produce significant effects–thus reported as "anticipatory"–even when a small proportion of looks to target is obtained. For instance, in the Kamide et al. (2003) study, the contrast between the two conditions was significant even when only 7 and 10% of the saccades have been made to the target, thus, during time slots at which 93 and 90% of all other saccades were directed elsewhere in the ersatz scenes. Notice also that because the scenes were impoverished and contained only about four to six potential targets (with a smaller number of potential targets at every new target fixated), reported effects of anticipatory eye movements often occur with proportions that are below chance, even if we were to assume that all objects are equipotential in terms of salience.

The ANOVAs and LME analyses took into account 31 bins of 50 ms each (up to 1550 ms), which corresponded to the maximum length of the noun complement. For the analysis of cumulative saccades (see **Figure 6**), there was a main effect of motion type ($F1(2, 60) = 59.85$, $p < 0.0001$, $\eta_p^2 = 66.6$, observed power = 1); $F2(2, 32) = 4.49$, $p = 0.019$, $\eta_p^2 = 21.9$, observed power = 0.727), a main effect of verb type only for participants ($F1(1, 30) = 46.62$, $p < 0.0001$, $\eta_p^2 = 60.8$, observed power = 1; $F2(1, 16) = 0.01$, $p = 0.932$, $\eta_p^2 = 0.01$, observed power = 0.051), and a significant verb × motion interaction by participants only ($F1(2, 60) = 35.29$, $p < 0.0001$, $\eta_p^2 = 54$, observed power = 1; $F2(2, 32) = 0.41$, $p = 0.67$, $\eta_p^2 = 2.5$, observed power = 0.11). The only difference between these analyses and the previous types of analyses (SOT and number of fixations) was that of a main effect of verb type, as well as an interaction between verb type and motion type (by participants), mostly because cumulative fixations take into account verb effects over time, when fixations to the target object are close to 1 for the two toward conditions. Planned comparisons revealed that in the toward motion condition, the cumulative number of saccades to the target object was greater when the verbs were causatives than when they were perception/psychological, in the participants' analysis ($t1(30) = 7.26$, $p < 0.0001$; $t2(16) = 0.77$, $p = 0.45$). No other comparisons between verb types across motion conditions were significant. It is important to highlight the main finding stemming from this analysis: The number of saccades to the target occurring at noun offset (750 ms bin from verb onset) is relatively small ($M = 0.4$), even in the fastest causative-toward condition. It appears that this condition shows an earlier peak than the other conditions during the processing of the noun, although the difference is only significant when noun offset is considered in the analysis. This effect suggests that this condition is set apart early on, although–as in the SOT analysis–this is not indicative of an anticipatory effect to target.

For the LME analyses, motion type, verb type, and bin were entered as fixed effects, and participants and items as random predictors. This model was compared to a null model consisting of only random predictors and was found to provide a better fit to the data, $\chi^2(7) = 278.73$, $p < 0.001$. We found a main effect of motion type and bin, a significant interaction between motion type and verb type, as well as a marginally significant interaction between motion type and bin (see **Table 3**).

Planned comparisons revealed that there was a significant difference in the cumulative number of saccades to the target object between the causative toward and perception toward condition, $p = 0.006$, suggesting that participants made a greater number of saccades to the target object over time when the verbs were causatives and when the agent was moving toward the target object. Similarly, results also showed a significant difference between the toward and away motion type conditions, $p < 0.001$, the toward and neutral conditions, $p = 0.007$, but no statistically significant difference between the away and neutral conditions, $p = 0.999$.

GENERAL DISCUSSION

The present study addressed two fundamental questions on the relationship between linguistic and visual integration. The first was the nature of this integration, here operationalized as saccades to objects in a dynamic scene as a function of verb onset in the speech stream. The second was the architecture of the visual-linguistic interaction that affords our ability to "talk about what we see" as Macnamara (1978) put it. This second question bears on the nature of the representations that both systems compute which enables them to "talk" to each other. We assumed–following Jackendoff and others (e.g., Fodor, 1975;

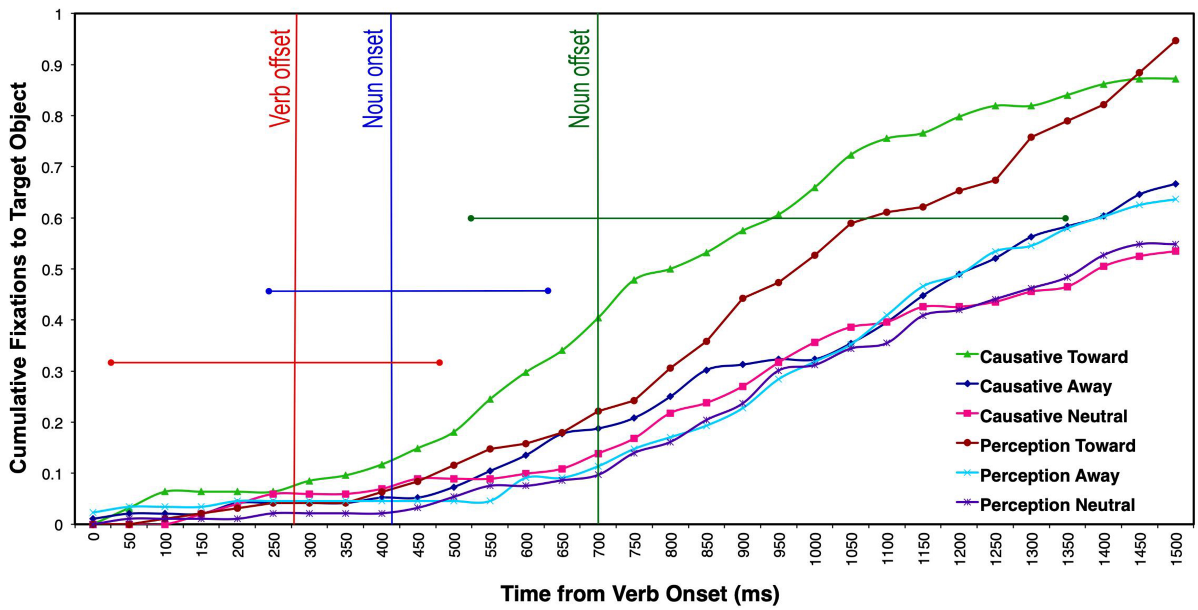


FIGURE 6 | Cumulative number of saccades launched to the target object after verb onset as a function of different motion and verb conditions. The vertical bars represent the mean verb offset (red), noun onset (blue), and noun offset (green) for the two linguistic conditions. The horizontal straight lines represent the length range of each of the three words.

TABLE 3 | Linear regression of cumulative number of saccades.

Predictor	β	SE β	*t*-value	95% CI of β	Null comparison
Constant	8.88	1.49	5.97	[0.60, 1.18]	
Motion type	1.56	4.99	3.12	[0.06, 0.25]	$\chi^2(1) = 5.58$, $p = 0.018$
Verb type	7.35	8.34	0.88	[−0.09, −0.24]	$\chi^2(1) = 0.76$, $p = 0.379$
Bin	1.61	6.57	2.45	[0.00, 0.03]	$\chi^2(1) = 5.99$, $p = 0.014$
Motion type × Verb type	−8.67	2.56	−3.40	[−0.14, −0.04]	$\chi^2(1) = 11.53$, $p < 0.001$
Motion type × Bin	3.52	1.97	1.79	[0.003, 0.007]	$\chi^2(1) = 3.21$, $p < 0.073$
Verb type × Bin	2.75	3.24	0.85	[0.004, 0.009]	$\chi^2(1) = 0.72$, $p = 0.395$
Motion type × Verb × Bin	−2.54	−3.94	−0.64	[−0.01, 0.01]	$\chi^2(1) = 0.41$, $p = 0.520$

Jackendoff, 1987, 2012) – that the interaction between language and vision requires amodal, rather than modality-specific codes. We begin our general discussion by addressing the effects we obtained with dynamic scenes and contrasting them with those obtained in other, static-scene visual world studies. We follow this with a discussion of the model we presented schematically in the Introduction, addressing in particular the nature of the interaction between language and vision and the representations they might deploy.

The Present Findings and Contrast With Static Visual World Studies

We made two main predictions. First, consistent with an interactive view of visual-linguistic architecture, we predicted that we would obtain faster saccades in the case of the more restrictive causative verb, than in the case of the perception/psychological verb. This prediction was consistent with those made in other studies (e.g., Altmann and Kamide, 1999; Staub et al., 2012) manipulating static scenes. Second, we predicted an interaction between verb type and context such that causatives would lead to faster saccades across all motion conditions. Moreover, we predicted anticipatory eye movements in the case of causatives which would signal an early interaction between verb meaning and scene. In contrast with these predictions, a lack of verb-type effect across motion conditions, and a lack of *anticipatory* verb effects would signal that viewers/listeners process linguistic and visual information independent of each other during initial stages, with a late interaction between the two systems.

Our findings seem to show a lack of early interaction. We obtained agent motion effects and a weak effect of verb type in the toward motion condition only, with causative verbs yielding marginally faster saccades to the target than perception verbs in the items analyses, as well as smaller number of saccades, and in the cumulative saccades over time. Moreover, we found no anticipatory effects in any of the conditions. Thus, hearing *crack* in a cooking context does not lead to anticipatory saccades to a highly related object *egg*, not even when the agent moves toward the object. This might suggest that background information computed from the scene (viz., the encoding of objects that populate a scene), or expectations about the unfolding event, do not influence initial verb-complement processing. In fact, it appears that eye movement processes are independent of linguistic interpretation when the scene is dynamic. That is, in dynamic scenes, attentional processes seem to lock onto relevant properties of the visual event–such as the *Agent* activity–without being influenced by properties of the linguistic stream, in particular by the selection of verb-complement *Theme* referents. This suggestion, however, should be seen with caution given that the lack of interaction between verb and motion type in the SOT analyses yielded low power.

We found that most of the saccades and brief fixations during the initial processing of the event were to the region of the agent (on average, this was a time window of 749 ms, going from the acoustic onset of the verb to the acoustic offset of the noun complement of the verb). These observations are in part consistent with previous studies on static scene processing without linguistic stimuli, which have found that focal attention is primarily directed toward human figures (Henderson and Ferreira, 2004; see also Yarbus, 1967, Ch. 7). This is also consistent with findings by Kamide et al. (2003) and Boland (2005) who showed that most post-verbal fixations were to the animate agents in the scene.

Our results appear to be at odds with previous studies that showed influence of visual/background context on linguistic processing and, in particular, they are at odds with results suggesting that eye movements to static scenes are locked into ongoing linguistic processes of verb-thematic assignment. We have in fact questioned those results, in our review, because they are often based on a small difference between conditions, or because they rely on a small proportion of fixations, or occur relatively late, post verb-complement offset. In our analyses of cumulative saccades, we do find an effect of verb type, with causatives leading to more saccades to target objects over time; however, this effect only occurs between verb offset and noun offset. But it is important to note that a significant effect of verb type here cannot be indicative of an anticipatory effect. Notice that data at that segment represent only an average of one tenth of a saccade. As we pointed out in our review of the literature, we could not argue that there is an anticipatory effect with such a small proportion of data. Thus, although it is possible that an effect of verb restriction begins to appear early on during verb processing, the effect is restricted by a very limited amount of data, and only when one does not take latencies or number of fixations into account.

The small verb type effect found in the toward condition may be better explained as a late, *confirmatory* effect, that is, a late integration between verb and scene information. This is so because eye movements are triggered by post-verbal information, and occur only when agent motion is indicative of which object the agent is about to interact with. Recall that agent motion disambiguation (toward or away from target object) is synchronized with verb onset in the sentence and spans a window of time of about 234 ms (7 frames). The onset of agent motion may be rapidly combined with verb semantics, but surprisingly not before the noun information is available. When the agent is moving toward a particular object, eye movements are drawn to that location, with computed verb-thematic information being used to further *confirm* the potential saccade-landing site.

It is possible, however, to take this confirmatory effect as representing a true interaction between agent-directed motion and verb-*thematic* (thus, by hypothesis, *linguistic*) information. In our contrast between verb classes, we hypothesized that the verb's thematic properties, which for causative verbs require a particular object to undergo a change of state (a *Patient* or *Theme*; see Levin and Rappaport Hovav, 2005), would draw attention to a potential *Theme* in the event. We predicted that the same effect would not be found in the case of the perception/psychological class because, contrary to causatives, the *Experiencers* ("agents") of perception/psychological verbs are the very entities that supposedly undergo a "change of state" (or that experience an object). Dissociations between agentives and *Experiencer* verbs have been found in Alzheimer's and aphasia patients (Piñango, 2006; Manouilidou et al., 2009), and the relative difficulty of *Experiencer* verbs has been attributed to their non-canonical thematic structure (no *Agent* role). Thus, the predicted difference between verbs in these two conditions (*Theme* that undergoes a change of state by the *Agent*; and *Theme* that causes a change of state in the *Experiencer*) should have contributed to *enhance* the differences in eye-movement behavior, if thematic roles were constraining referents in the visual context. But this was not the case in two of the motion conditions. Even if what we observed in the causative-toward condition, then, is an effect of typically *linguistic* representations interacting with information computed about the scene, they represent late effects. As we will discuss, in the context of the model presented in the next section, the conceptual representation of verbs might rapidly interact with conceptual representation of scene information to direct gaze to objects.

A key component of our investigation is the use of dynamic scenes and the manipulation of visual/motion context. But this manipulation introduces variables that may hinder a direct comparison with studies employing still pictures and ersatz scenes. For instance, dynamic scenes might be more cognitively taxing, obfuscating an otherwise early interaction between visual context and linguistic processing. Andersson et al. (2011) demonstrated that when presented with highly complex static scenes coupled with fast-paced sentences containing four nouns referring to objects in the scenes, participants have difficulty keeping track of referents, yielding saccade latencies in the magnitude of 2500 ms post-noun onset. Our scenes, however, are not "hoarding" scenes as those employed by Andersson et al. (2011), our sentences were recorded at a normal pace, and referred primarily to only one object in the scene (internal argument of the verb), other than the agent, in the relevant contrast. In addition, as we discussed above, our verb contrasts and the toward motion condition should have contributed to *enhance* anticipatory effects.

It is also possible that agent motion in the scene grabs overt attention, even when the verb might be covertly directing attention to its related object–thus reflecting an early interaction that is not manifested in early saccades. We have no direct evidence that subjects might be withholding a saccade in spite of an attention switch to the target object. But we contend that fixations on moving agents were not simply a function of low-level motion because fixations remained on the agents for up to 2 s after the onset of motion, suggesting that attention to agents is mostly goal-directed instead of stimulus-driven (Hillstrom and Yantis, 1994). In addition, results from our laboratory, in which we employed a change blindness task with dissolving objects during realistic dynamic scenes, demonstrate a similar effect: the sudden motion of the dissolving target object did not capture attention, which is initially directed to the agent of the event (van de Velde, 2008). Related inattention blindness effects in dynamic scenes, but with human targets, have been reported (Simon and Chabris, 1999). Thus, it appears that the effects we obtained in the present study cannot be simply attributed to motion in the scene preventing attention to verb-related objects. And even if we were to attribute delayed saccades to the object due to agent motion, a typical linguistic by visual processes interaction would allow for verb effects to be obtained *consistently* within scene types, which was not the case in the present experiment.

As observed by a reviewer, one other factor may have potentially delayed saccades to targets contributing to our lack of anticipatory effects, even in the most constraining causative-toward condition: the post-experimental recall task, which could have promoted memorization rather than rapid comprehension of sentences and scenes. We contend, however, that our task promotes comprehension (scene gist, sentence proposition), which is what participants later use to recognize material in the post-experiment task. Other studies have encouraged subjects to look at objects (e.g., in Andersson et al., 2011) with questions at every trial. Ito et al. (2018), requested subjects to keep a list of words in mind while performing the visual world task (a cognitive load manipulation) and obtained different effects compared to a no-load condition. However, our post-experimental task is a simple cued recall task of sentences and scenes similar to asking subjects questions after a trial. We reiterate that in our causative-toward condition subjects have both verb and motion information to anticipate the target object.

On the flip side, anticipatory effects may occur primarily in static and ersatz scenes–thus they might not be the *norm* of language use across visual contexts. It is possible that in situations where the scene is static (for instance, dishes on a dinner table or objects on a shelf) the viewer/listener might be more sensitive to linguistic information, even anticipating the nature of referents.[8] The same applies to goal-directed tasks (e.g., Tanenhaus et al., 1995). It is to these situations that the studies we reviewed best apply. But their use to generalize about language use in context tout court can be challenged by the present results.

In summary, the results of our experiment suggest that visual attention and linguistic processing may be computed independently and in parallel, in the construction of *dynamic* event representations. It seems that the processing of dynamic naturalistic scenes and sentences occurs without visual attention being initially influenced by the nature of the linguistic stream, and without representations built from visual context influencing the early selection of linguistic referents *by necessity*. This apparent decoupling of linguistic and visual processes during the early moments of linguistic and visual-scene perception may be one further indication that the two systems are modular, rather than interactive, with interaction occurring at a later stage.

Where and *How* Do (the Products of) Vision and Language Interact?

A second more fundamental question that the present article addresses is the nature of the architecture of the visual-linguistic system and how their respective representations are combined. There seems to be little doubt in cognitive science that information computed by visual and linguistic systems *should be* integrated at some processing level and that this integration needs to rely on a common representational code (Fodor, 1975, 2008; Macnamara, 1978; Jackendoff, 1987, 2012). The assumption of a common representational code is not exclusive of symbolic approaches to cognition, as in the cited works, but a characteristic of highly interactive models as well (McClelland et al., 2014). The model we propose, however, takes linguistic and visual processes to share information at a higher-level conceptual system, with their early inputs being encapsulated.

A modular encapsulated system in a symbolic cognitive architecture is sensitive to *formal* properties of the information it computes (Fodor, 2000). In our model, these formal properties include, among others, word and sentence structure and argument/thematic properties of verbs, on the linguistic input, and feature combination, scene layout, and token object discrimination, in the visual input. These input systems are tuned to *different natural kinds*, and they appear to operate on different formal properties but to produce outputs that might serve the interpretation of events relying on common predicates.

One of the most promising attempts to conceive these common predicates that serve the interaction between linguistic and visual representations has been that of Jackendoff (see, in particular, Jackendoff, 1987, 2012). The model we present in **Figure 7** proposes a similar view of the visual and linguistic systems (For a more detailed description of the workings of the model, see **Supplementary Figure S2**).

The model takes the early inspection of a scene concomitantly with the analysis of linguistic input to be computed independently and in parallel by the two systems. These two systems bind scene and sentence constituents into predicate-like structures. More specifically, relations between objects in the scene and unfolding events are represented in a language that is common to the *output* of both visual and linguistic systems. In the example in **Figure 7**, the computation of propositions is couched in a more neutral predicate-logic notation to exemplify how representations of objects and their relations obtained from the unfolding linguistic and visual domains might be combined. In Jackendoff's (1987, 1990) more detailed notation, conceptual structures related to *States* and *Events* (as well as other ontological categories such as *Thing* and *Place*) might be continuously computed from scene information, thus forming a framework for the interpretation of unfolding linguistic

[8] We thank a reviewer for stressing this point.

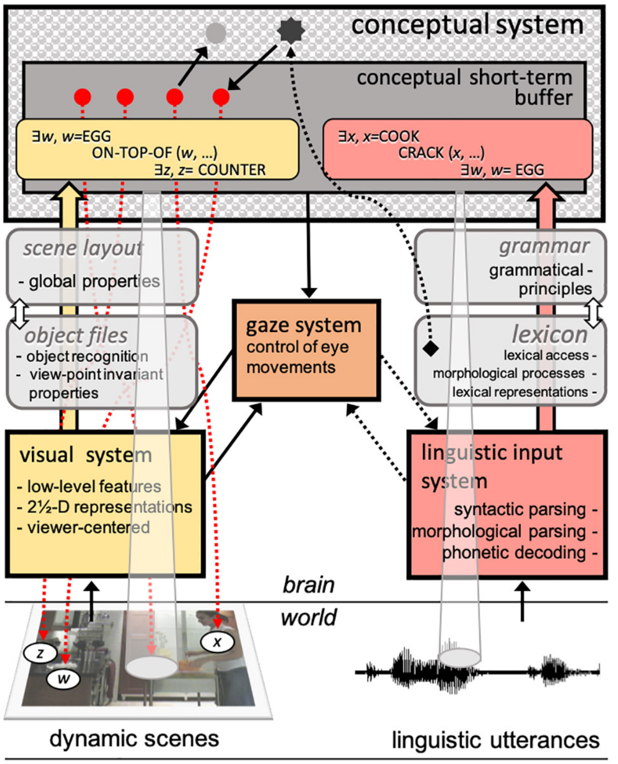


FIGURE 7 | The workings of the modular dynamic visual-linguistic interaction model. See **Supplementary Figure S2** for detailed description.

messages. These conceptual-structure representations computed from an event such as the one presented in **Figure 1** may consist of expressions such as those in (1).

(1) (a.) [$_{State}$ BE ([$_{Thing}$ WOMAN], [$_{Place}$ IN ([$_{Thing}$ KITCHEN])])]
(b.) [$_{State}$ BE ([$_{Thing}$ EGG], [$_{Place}$ ON ([$_{Thing}$ COUNTER])])]
(c.) [$_{Event}$ PUT ([$_{Thing}$ WOMAN ([$_{Event}$ FLOUR, [$_{Place}$ IN ([$_{Thing}$ BOWL])])])])]
(d.) [$_{Event}$ GO [$_{Thing}$ WOMAN [$_{Path}$LEFT]]])[9]

This proposal is also compatible with the idea that dynamic scene inspection and action require the binding and tracking of multiple elements within the "perceptual circle." These elements are represented as visual predicates computed early on about indexed (or FINSTed) scene constituents (see, e.g., Fodor and Pylyshyn, 2015; Pylyshyn, 2018).

Besides being explicit about the types of representations that might constitute the interaction between linguistic and visual processes, an advantage of the present proposal over connectionist models mentioned above, is that the codes that constitute conceptual structure (and in particular those temporarily placed in the CSTB) are computable; that is, the unfolding processes are carried over as a function of the formal properties of the expressions. The expressions are also compositional–so that the propositions that they express are indeed a function of their constituent elements and their structure. Simply put, connectionist networks need to account for how the nodes that stand for, say, WOMAN, CRACK, COUNTER, and EGGS, bear the relation that they do such that the sentence means that the woman will crack the eggs that are on the counter. A similar point can be made about the concepts computed from the scene: they need to be bound into events. Without postulating *structural relations* between the "activated" elements, there is no saying on what they mean together (see Fodor and Pylyshyn, 1988).

Sentence processing during dynamic events occurs in contexts that thus might be rich in conceptual representations computed *from* scene inspection–with sentences contributing to attention focusing on (and "indexing" of) what is happening in the scene. In the case of the study reported here, attention is directed toward the interaction between *Agent* and affected object or between *Agent/Experiencer* and the causer of psychological state (both referents of *Themes* in the sentences). As can be seen in **Figure 7**, we assume that the semantic representation of the sentence also takes these same conceptual-structure codes, thus contributing content to the ongoing conceptual representation of the event that is seen and heard concomitantly. In sum, the interaction between predicate structures computed from scene and sentence representations are computed in parallel and integrated in real time based on the products of their respective input systems.

It is difficult to say whether the effects we obtained are modular in the now classical sense (Fodor, 1983) or whether they represent a type of encapsulation typical of even higher cognitive systems (Barrett and Kurzban, 2006; Chomsky, 2018). The classical type of modularity system encompasses only mandatory perceptual computations, which are domain-specific, and are not influenced by other input systems or by general cognitive processes such as those of long-term memory, expectations, and desires. The higher-cognitive type of modular system exhibits some of the properties of input modules, but is primarily defined by the type of knowledge domain it operates on–akin to Chomsky's (2018) view of modules as central systems. Thus, even logical inferences might be *modular* if they are, as Barrett and Kurzban (2006) suggest, "construed in terms of the formal properties of information that render it processable by some computational procedure" (p. 634). In the present case, it is possible to assume that visual contextual information influences on language are "domain-specific" in this sense, if language-vision interactions are computed over a common code at a higher level than their classical input modules. It could be the case that vision and language have *evolved* to work together in action and communication, and thus, they might operate with representations from both classical input domains. If this is the case, then effects of context on parsing decisions (such as those reported by Tanenhaus et al., 1995; Spivey et al., 2002) or putative effects of context on thematic-role processing, as

[9]In this example, we adopt a language-like representation following Jackendoff's (1990) notation as a convenient way to represent the potential predicate structures computed from the event. But we do not assume that the predicates themselves are further decomposed, nor do we take the ontological categories–such as *Thing*, *Event*, etc.–to be necessarily the ones we encode from scenes or sentence meanings, although we see heuristic value in them.

measured by eye-movement behavior, could be manifestations of a common domain of knowledge, a modular higher-cognitive system in the sense proposed by Barrett and Kurzban (2006) and Chomsky (2018). It is difficult to tease apart these proposals for modularity but they do constitute alternative interpretations for some now established effects of context in language and for the effects we found.

All studies employing the VWP with static scenes stress how fast the interaction between visual and linguistic representations might be, given the anticipatory or early post-verbal effects commonly found. Contrary to static scenes, dynamic events as the ones employed in the present study, rely on representations computed from both visual and linguistic inputs being continuously updated, that is, they need to allow for dynamic interaction of information in working memory. In our model, these computations are occurring at the CSTB (akin to Potter, 1999; see also Potter, 2018). The system accesses long-term memory representations of words and objects and builds conceptual structures compatible with both visual and linguistic predicates. We suggest, then, that the locus of the *effects* of visual contextual influence in sentence processing is post-perceptual, that is, when both visual and linguistic *outputs* have reached the conceptual buffer and expressions about the unfolding event are being built.

Given that access to conceptual information about objects and scenes may occur within 100–200 ms of scene onset (Potter and Faulconer, 1975; Thorpe et al., 1996; see also **Supplementary Figure S2**), it would be expected that information about the scene would guide the gaze to the appropriate referents as sentences unfold–and in particular, as verbs make certain objects in the naturalistic scenes potentially more prominent for further processing. If so, objects of causative sentences and their referents in the world would have an advantage over objects of perception verbs. Clearly, in our study, eye movements are controlled more by what the visual context "says" about the event than by what the sentence says, and very little by the interaction between the two systems. Although most research on scene gist processing has been done with static scenes (see Henderson and Hollingworth, 1999), eye movement studies on scene processing suggest that while the gist is obtained rapidly, consolidation of scene details continues and requires possibly indexing and serial visual routines to integrate information (Pylyshyn, 2003). The processing of the category or the gist of a scene may rely on determining the meaning or category of one of its constituent objects; the initial representation of objects and scene, however, need not be conceptual: Henderson and Hollingworth (1999) as well as Pylyshyn (2003, 2007) point to the structural or even "pre-semantic" nature of scene perception in vision. What happens after the initial analysis of the scene requires further computations–processes over conceptual-structure representations– which are likely to take into account sources of information such as the products of linguistic input. We take these high-level computations and the context effects that they engender to occur post-perceptually, relying on the structural analyses provided by both language and visual input systems.

CONCLUSION

We have raised concerns about the generalizability of studies that support a fully interactive–rather than a modular–view of functional architecture using impoverished static rather than realistic and dynamic scenes. One of our primary concerns is that the claims made in support of an interactive–so called *incremental*–linguistic system rely on stimuli that do not necessarily represent the use of language in dynamic visual contexts. Stimulus variables such as scene complexity and motion call into question the conclusion that the linguistic system takes contextual information into consideration at a very early stage, with linguistic input analysis being sensitive to information available to the visual system (Tanenhaus et al., 1995). Given the potential usefulness of the visual-world paradigm for understanding how different perceptual and cognitive systems make their representations available to each other, it is important to consider all possible alternative *loci* of influence or interface between the representations computed by different input and cognitive systems.

We suggest, then, that the locus of the previously observed visual context effects on sentence processing may actually be *after* the initial analysis of linguistic tokens, where alternative sentence parses or interpretations may be selected for further processing or may be re-analyzed according with a particular visual context representation. In the present study, eye movements to referents of verb arguments occur after the *offset* of the noun complement (ranging from 140 ms in the causative-toward condition to 700 ms in the perception-neutral condition). At that point verb-thematic properties across contexts have been processed, well in advance of eye movements to verb-related objects. Visual context may help guide attention to objects in the scene only when the *interpretation* of the critical verb phrase is under way. Agents of dynamic events, however, appear to grab the focal attention of the viewer/hearer, thus making eye movements to what is seen initially insensitive to what is heard–in particular without the seemingly mandatory verb-structural and thematic effects found in studies with static scenes (e.g., Altmann and Kamide, 1999; Knoeferle et al., 2005; Staub et al., 2012). In our study, "what was heard" was not ignored, as participants were able to interpret/encode the sentences despite not tracking word-referents on the screen continuously (see also Andersson et al., 2011) and without being sensitive to verb-thematic properties in two of the motion conditions. That is, sentences and scenes were independently processed and information about both was integrated at post-perceptual stages; in our proposal, the post-perceptual interaction between the two input systems occurs at a conceptual buffer and takes the form of common-code conceptual, predicate-like structures.

In conclusion, what is striking about the effect we obtained–an insensitivity to verb distinctions when agents do not engage objects–is that it is commonly believed that our attention is usually tied to the products of our linguistic processes, in particular in visual contexts (Tanenhaus et al., 1995; Altmann and Kamide, 1999; Spivey et al., 2002). The popularity of the visual world technique (see Huettig et al., 2011) attests to its perceived usefulness to investigate linguistic processes and their interaction with visual context. But we question its effects insofar as impoverished scenes are employed. We found that the process of understanding a dynamic visual event and understanding a sentence describing the event appear to be largely independent of each other during the initial processing of visual and linguistic input. The caveat is that the two processes–visual and linguistic–interact with each other when, besides linguistic cues, agents disambiguate the nature of their actions. When this happens–for instance, when agents walk toward a given object–then verbs that are supposed to "select" for these objects trigger faster and a smaller number of saccades to these objects than neutral verbs do. However, even in these cases, when causative verbs seem to constrain the domain of reference to one object and when the agent reaches toward that object, saccades are *not* anticipatory, thus challenging results supporting interactive models of language processing. Further, from a methodological standpoint, it seems clear that the use of dynamic rather than static scenes constitutes a more ecologically valid method in the study of the potential influence of visual context on language comprehension, thus representing an advance in the investigation of the interaction (or lack thereof) between these key cognitive and perceptual systems. We believe we found support for the independence–or modularity–of linguistic and visual processes employing a task that better exemplifies realistic uses of language in dynamic scenes. We have proposed that the two systems interact only at a central, conceptual system that operates over the *outputs* of vision and language input systems, and relies on a common, propositional code. Advancing Macnamara (1978) quest for understanding how we talk about what we see (or how we understand what we hear and see concomitantly), we have proposed a model that takes visual and linguistic predicate-argument relations to be the basis of language use in visual contexts.

ETHICS STATEMENT

This study was conducted in accordance with the recommendations of the Concordia University Human Research Ethics Committee and was approved by this committee. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

AUTHOR CONTRIBUTIONS

RA is the main author of the manuscript, the principal responsible for the conception and design of the study, production of materials, theoretical background, and general discussion. JD conducted the experiment, and contributed to the statistical analyses and to the manuscript. CA performed the main statistical analyses and contributed to the final manuscript. MG contributed to the conception and design of the study, and to the production of materials.

FUNDING

This research was supported by grants from the *Fonds Québécois de la Recherche sur la Société et la Culture*, from the Social Sciences and Humanities Research Council, and from the Natural Sciences and Engineering Research Council (NSERC) to RA, by an NSERC grant to MG, and by an NSERC graduate fellowship to JD.

ACKNOWLEDGMENTS

We are indebted to Zinnia Madon for helping with the preparation of the stimuli and data collection, and to Caroline de Almeida for producing and editing the video clips. We thank Fernanda Ferreira, Sandy Pollatsek, Molly Potter, Tonje Persson, Levi Riven, Norman Segalowitz, and the two reviewers for comments on earlier drafts. We are also indebted to Vincent Porretta for discussion on statistical analyses. Our colleague passed away before seeing the latest version of the present article. We hope it does justice to his outstanding contributions to visual perception and attention.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02162/full#supplementary-material

REFERENCES

Altmann, G. T. M., and Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition* 73, 247–264. doi: 10.1016/s0010-0277(99)00059-1

Altmann, G. T. M., and Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: linking anticipatory (and other) eye movements to linguistic processing. *J. Mem. Lang.* 57, 502–518. doi: 10.1016/j.jml.2006.12.004

Andersson, R., Ferreira, F., and Henderson, J. M. (2011). I see what you're saying: the integration of complex speech and scenes during language comprehension. *Acta Psychol.* 137, 208–216. doi: 10.1016/j.actpsy.2011.01.007

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Baddeley, A. (2012). Working memory: theories, models, and controversies. *Annu. Rev. Psychol.* 63, 1–29. doi: 10.1146/annurev-psych-120710-100422

Barrett, H. C., and Kurzban, R. (2006). Modularity in cognition: framing the debate. *Psychol. Rev.* 113, 628–647.

Bates, D., Maechler, M., and Bolker, B. (2013). *Lme4: Linear Mixed Effects Models Using S4 Classes, R Package Version 0.9999099-2*. Available at: http://CRAN.R-project.org/package=lme4

Boland, J. E. (2005). Visual arguments. *Cognition* 95, 237–274. doi: 10.1016/j.cognition.2004.01.008

Chomsky, N. (2018). "Two notions of modularity," in *On Concepts, Modules, and Language: Cognitive Science at its Core*, eds R. G. de Almeida, and L. R. Gleitman, (New York, NY: Oxford University Press), 25–40.

Coco, M. I., Keller, F., and Malcolm, G. L. (2016). Anticipation in real-world scenes: the role of visual context and visual memory. *Cogn. Sci.* 40, 1995–2024. doi: 10.1111/cogs.12313

Cohen, J. D., MacWhinney, B., Flatt, M., and Provost, J. (1993). PsyScope: a new graphic interactive environment for designing psychology experiments. *Behav. Res. Methods Instrum. Comput.* 25, 257–271. doi: 10.3758/bf03204507

Coltheart, M. (1981). The MRC psycholinguistic database. *Q. J. Exp. Psychol.* 33A, 497–505. doi: 10.1080/14640748108400805

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language. *Cogn. Psychol.*<refvol>6, 84–107. doi: 10.1016/0010-0285(74)90005-x

de Almeida, R. G., and Antal, C. (in press). "How can semantics avoid the troubles with the analytic/synthetic distinction?," in *Conceptual Structures: Linguistic, Philosophical, and Psychological Perspectives*, ed. T. Gamerschlag, (New York, NY: Springer).

de Almeida, R. G., and Lepore, E. (2018). "Semantics for a module," in *On Concepts, Modules, and Language: Cognitive Science at its Core*, eds R. G. de Almeida, and L. R. Gleitman, (New York, NY: Oxford University Press), 113–138.

de Almeida, R. G., and Manouilidou, C. (2015). "The study of verbs in cognitive science," in *Cognitive Science Perspectives on Verb Representation and Processing*, eds R. G. de Almeida, and C. Manouilidou, (New York, NY: Springer).

Fischer, B., and Ramsperger, E. (1984). Human express saccades: extremely short reaction times of goal directed eye movements. *Exp. Brain Res.* 57, 191–195.

Fodor, J. A. (1975). *The Language of Thought*. Cambridge, MA: Harvard University Press.

Fodor, J. A. (1983). *The Modularity of Mind*. Cambridge, MA: MIT Press.

Fodor, J. A. (2000). *The Mind Doesn't Work that Way*. Cambridge, MA: MIT Press.

Fodor, J. A. (2008). *LOT 2: The Language of Thought Revisited*. New York, NY: Oxford University Press.

Fodor, J. A., and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition* 28, 3–71. doi: 10.1016/0010-0277(88)90031-5

Fodor, J. A., and Pylyshyn, Z. W. (2015). *Minds Without Meanings: An Essay on the Content of Concepts*. Cambridge, MA: MIT Press:.

Frazier, L. (1988). "Grammar and language processing," in *Linguistics: The Cambridge survey, Vol. 2. Linguistic theory: Extensions and implications*, ed. F. J. Newmeyer, (New York, NY: Cambridge University Press), 15–34.

Henderson, J. M., and Ferreira, F. (2004). "Scene perception for psycholinguists," in *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*, eds J. Henderson, and F. Ferreira, (New York, NY: Psychology Press), 1–58.

Henderson, J. M., and Hollingworth, A. (1999). High-level scene perception. *Annu. Rev. Psychol.* 50, 243–271. doi: 10.1146/annurev.psych.50.1.243

Hillstrom, A. P., and Yantis, S. (1994). Visual motion and attention capture. *Percept. Psychophys.* 55, 399–411.

Huettig, F., Rommers, J., and Meyer, A. S. (2011). Using the visual world paradigm to study language processing: a review and critical evaluation. *Acta Psychol.* 137, 151–171. doi: 10.1016/j.actpsy.2010.11.003

Ito, A., Corley, M., and Pickering, M. J. (2018). A cognitive load delays predictive eye movements similarly during L1 and L2 comprehension. *Biling. Lang. Cogn.* 21, 251–264. doi: 10.1017/s1366728917000050

Jackendoff, R. (1987). *Consciousness and the Computational Mind*. Cambridge, MA: MIT Press.

Jackendoff, R. (1990). *Semantic Structures*. Cambridge, MA: MIT Press.

Jackendoff, R. (2012). Language as a source of evidence for theories of spatial representation. *Perception* 41, 1128–1152. doi: 10.1068/p7271

Kamide, Y., Altmann, G. T. M., and Haywood, S. (2003). Prediction and thematic information in incremental sentence processing: evidence from anticipatory eye-movements. *J. Mem. Lang.* 49, 133–156. doi: 10.1016/s0749-596x(03)00023-8

Knoeferle, P., and Crocker, M. W. (2006). The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking. *Cogn. Sci.* 30, 481–529. doi: 10.1207/s15516709cog0000_65

Knoeferle, P., Crocker, M. W., Scheepers, C., and Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: evidence from eye-movements in depicted events. *Cognition* 95, 95–127. doi: 10.1016/j.cognition.2004.03.002

Kucera, H., and Francis, W. N. (1967). *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press.

Land, M. F. (2009). Vision, eye movements, and natural behavior. *Vis. Neurosci.* 26, 51–62. doi: 10.1017/S0952523808080899

Lenth, R., Singmann, H., Love, J., Buerkner, P., and Herve, M. (2018). *Package "Emmeans", R Package Version 4.0-3*. Available at: https://cran.r-project.org/web/packages/emmeans/index.html

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.

Levin, B., and Rappaport Hovav, M. (2005). *Argument Realization*. Cambridge, MA: Cambridge University Press.

Macnamara, J. (1978). *How Can We Talk About What We See?* Montreal: McGill University.

Manouilidou, C., de Almeida, R. G., Schwartz, G., and Nair, N. P. V. (2009). Thematic roles in Alzheimer's disease: hierarchy violations in psychological predicates. *J. Neurolinguistics* 22, 167–186. doi: 10.1016/j.jneuroling.2008.10.002

Mayberry, M. R., Crocker, M. W., and Knoeferle, P. (2009). Learning to attend: a connectionist model of situated language comprehension. *Cogn. Sci.* 33, 449–496. doi: 10.1111/j.1551-6709.2009.01019.x

McClelland, J. L., Mirman, D., Bolger, D. J., and Khaitan, P. (2014). Interactive activation and mutual constraint satisfaction in perception and cognition. *Cogn. Sci.* 38, 1139–1189. doi: 10.1111/cogs.12146

McClelland, J. L., Rumelhart, D. E., and Hinton, G. E. (1986). "The appeal of parallel distributed processing," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, eds D. E. Rumelhart, J. L. McClelland, and the PDP Research Group (Cambridge, MA: MIT Press) 1, 3–44.

Münte, T. F., Schiltz, K., and Kutas, M. (1998). When temporal terms belie conceptual order. *Nature* 395, 71–73. doi: 10.1038/25731

Piñango, M. M. (2006). "Thematic roles as event structure relations," in *Semantic Role Universals and Argument Linking: Theoretical, Typological, and Psycholinguistic Perspectives*, eds I. Bornkessel, M. Schlesewsky, and A. Friederici, (Berlin: Mouton de Gruyter), 303–353.

Potter, M. C. (1999). "Understanding sentences and scenes: the role of conceptual short-term memory," in *Fleeting Memories: Cognition of Brief Visual Stimuli*, ed. V. Coltheart, (Cambridge, MA: MIT Press), 13–46.

Potter, M. C. (2018). "The immediacy of conceptual processing," in *On concepts, modules, and language: Cognitive science at its core*, eds R. G. de Almeida, and L. R. Gleitman, (New York, NY: Oxford University Press), 239–248.

Potter, M. C., and Faulconer, B. A. (1975). Time to understand pictures and words. *Nature* 253, 437–438. doi: 10.1038/253437a0

Pylyshyn, Z. W. (2003). *Seeing and Visualizing: It's not What you Think*. Cambridge, MA: MIT Press.

Pylyshyn, Z. W. (2007). *Things and Places: How the Mind Connects With the World*. Cambridge, MA: MIT Press.

Pylyshyn, Z. W. (2018). "Scientific theories and Fodorian exceptionalism," in *On Concepts, Modules, and Language: Cognitive Science at its Core*, eds R. G. de Almeida, and L. R. Gleitman, (New York, NY: Oxford University Press), 191–208.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing. R-Foundation for Statistical Computing*. Available at: http://www.R-project.org/

Reichle, E. D., Rayner, K., and Pollatsek, A. (2003). The E-Z reader model of eye-movement control in reading: comparisons to other models. *Behav. Brain Sci.* 26, 445–476. doi: 10.1017/s0140525x03000104

Simon, D. J., and Chabris, C. F. (1999). Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception* 28, 1059–1074. doi: 10.1068/p2952

Smith, A. C., Monaghan, P., and Huettig, F. (2017). The multimodal nature of spoken word processing in the visual world: testing the predictions of alternative models of multimodal integration. *J. Mem. Lang.* 93, 276–303. doi: 10.1016/j.jml.2016.08.005

Spivey, M. J., Tanenhaus, M. K., Eberhard, K. M., and Sedivy, J. C. (2002). Eye movements and spoken language comprehension: effects of visual context on syntactic ambiguity resolution. *Cogn. Psychol.* 45, 447–481. doi: 10.1016/s0010-0285(02)00503-0

Staub, A., Abbott, M., and Bogartz, R. S. (2012). Linguistically guided anticipatory eye movements in scene viewing. *Vis. Cogn.* 20, 922–946. doi: 10.1080/13506285.2012.715599

Tanenhaus, M. K., and Spivey-Knowlton, M. J. (1996). Eye-tracking. *Lang. Cogn. Process.* 11, 583–588.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 1632–1634. doi: 10.1126/science.7777863

Thorpe, S. J., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature* 381, 520–522. doi: 10.1038/381520a0

van de Velde, C. (2008). *Tracking Eye Movements to Uncover the Nature of Visual-Linguistic Interaction in Dynamic and Static Scenes*. Montreal: Concordia University. Ph.D. dissertation.

Yarbus, A. L. (1967). *Eye Movements and Vision*. New York, NY: Plenum Press.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Communication

ORIGINAL RESEARCH
published: 01 November 2019
doi: 10.3389/fcomm.2019.00054

Check for updates

The Morphophonology of Intraword Codeswitching: Representation and Processing

Sara Stefanich[1], Jennifer Cabrelli[2], Dustin Hilderman[3] and John Archibald[3]*

[1] Department of Spanish and Portuguese, Northwestern University, Evanston, IL, United States, [2] Department of Hispanic and Italian Studies, University of Illinois at Chicago, Chicago, IL, United States, [3] Department of Linguistics, University of Victoria, Victoria, BC, Canada

This paper serves as a critical discussion of the phenomenon of intraword code-switching (ICS), or the combining of elements (e. g., a root and an affix) from different languages within a single word. Extensive research over the last four decades (Poplack, 1988; Myers-Scotton, 2000; MacSwan, 2014) has revealed CS to be a rule-governed speech practice. While interword CS is widely attested, intraword code-switching has been argued to be impossible (Poplack, 1980; Bandi-Rao and den Dikken, 2014; MacSwan and Colina, 2014). However, ICS has recently been documented in language pairs ranging from English/Norwegian (Alexiadou et al., 2015) to Nahuatl/Spanish (MacSwan, 1999) to Greek/German (Alexiadou, 2017), and is a robust phenomenon. We review the foundational research on ICS, followed by an examination of the phenomenon from the perspectives of knowledge and skill. First, we examine intraword CS as part of a bilingual's I-language to determine the morphological and phonological restrictions on the phenomenon. We operationalize these restrictions within a Distributed Morphology (DM) framework (e.g., Halle and Marantz, 1994) in which the traditional lexicon is split into three lists. List 1 contains lexical roots and grammatical features or feature bundles, while Lists 2 and 3 detail instructions for phonological realization (i.e., rules for Vocabulary Insertion) and semantic interpretation, respectively. Here we probe the question of whether words which have morphological mixing also have phonological mixing. Second, building on the DM machinery, we present an account for intraword CS in performance via the modular cognitive performance framework of MOGUL (Sharwood Smith and Truscott, 2014). This modular architecture assumes (a) that lexical items are constituted by chains of representations and (b) that extra-linguistic cognitive mechanisms (e.g., goals, executive control) play a role in ICS (Green and Abutalebi, 2013). ICS is licensed by a bilingual mode of communication (following Grosjean, 2001) where the act of CS itself serves an illocutionary goal; it is the real-world context which triggers the complex CS system. Thus, viewing intraword CS as an I-language and an E-language phenomenon provides an explanatory model of the dynamic knowing that and knowing how which is manifest in the phenomenon of ICS.

Keywords: codeswitching, intraword codeswitching, Distributed Morphology, MOGUL, morphophonology

OPEN ACCESS

Edited by:
Gary Libben,
Brock University, Canada

Reviewed by:
Terje Lohndal,
Norwegian University of Science and Technology, Norway
Artemis Alexiadou,
Humboldt University of Berlin, Germany
Anna Hatzidaki,
National and Kapodistrian University of Athens, Greece

***Correspondence:**
John Archibald
johnarch@uvic.ca

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *31 October 2018*
Accepted: *30 September 2019*
Published: *01 November 2019*

Citation:
Stefanich S, Cabrelli J, Hilderman D and Archibald J (2019) The Morphophonology of Intraword Codeswitching: Representation and Processing. Front. Commun. 4:54. doi: 10.3389/fcomm.2019.00054

INTRODUCTION

It is uncontroversial to note that bilinguals sometimes switch languages within a conversation and even within a sentence. In this paper, we seek to explore an even smaller domain of the linguistically-mixed *word*. When we use the word *bilingual*, we intend a very broad interpretation which runs the gamut from classroom second language learners to professional simultaneous interpreters. In a predominantly multilingual world, there are many conversations each day which involve people who know more than one language. They may occur within a multilingual family, or friend group, or society. There are two basic facts which underlie this seemingly effortless performance:

(A) bilinguals are normally very good at suppressing the production of a language which is not the language of the current environment, and
(B) the goal of the speaker is to have each utterance in a conversation successfully communicate an intended meaning, and will recruit all available linguistic resources to do so

Given these assumptions, it is interesting for the linguist and the psycholinguist to probe what underlies this phenomenon known as codeswitching, as, on the surface, it would appear that codeswitching might appear to violate both of the above axioms (given that one language is not suppressed, and that perhaps some of the interlocutors do not speak both languages). Research over the past 50 years has revealed that switching languages is not a sign of impoverished linguistic ability, or low proficiency, but rather is a complex performative dance which involves exchanging information, marking solidarity, and revealing identity (Poplack, 1980). However, like many aspects of linguistic performance, the rules and patterns are not open to conscious inspection. Just as when our knee is itchy, we "simply" invoke the motoric commands to scratch (and could not articulate them), when we have a message to convey to a particular listener or group of listeners, we "simply" produce the utterance (and could not explicitly state the grammatical rules which generated that utterance). This is true of both monolingual and bilingual utterances.

Throughout this paper, we will adopt a narrative strategy of referring to two groups of bilingual speakers (one Spanish/English, one Norwegian/English) in a casual, conversational manner to help elucidate some of the technical constructs. There is a rich corpus of data from heritage Norwegian speakers in the United States (Johannessenn, 2015) which we will draw on frequently in this paper. Let us imagine a gathering of members of this group in someone's home to have coffee and cake, and to watch a sporting event. If we were a fly (or tape-recorder) on the wall, we would undoubtedly hear many utterances which had elements of both English and Norwegian. Some of these utterances are produced by a (hypothetical) member of the community whom we will call Gunnar. Perhaps when referring to a particular sandwich type being served, Gunnar will use the Norwegian word *Smørbrød* as this seems more appropriate than the English *sandwich*. But perhaps when a hockey game comes on the TV, Gunnar is reminded of the Lillehammer winter Olympics in 1994 and begins to talk about sports completely in Norwegian. Then this sports talk is interrupted by a phone call which proceeds entirely in English. After the phone call, when the discussion switches to the upcoming state elections, the discussion switches to a mixture of Norwegian and English as people talk of both American and Norwegian politicians. This mixture of languages may include language switches within a sentence (intrasentential), switches between sentences (intersentential), or our main focus here, switches within a word (intraword). Such examples of multilingual communication are very common. They occur as effortlessly and automatically as any monolingual conversation.

However, mixing languages within a single word is a seemingly small linguistic phenomenon with surprisingly far-reaching implications, and an interesting history. For one thing, in the literature on codeswitching, there are researchers who have denied that intraword codeswitching (ICS) is even possible. We will argue that it is a robust and widespread characteristic of multilingual speech, and propose mechanisms to account for both the knowledge and performance systems which generate these forms. Surveying examples presented in the literature, it is clear that morphological elements can be combined in a single word in a systematic manner, just like codeswitching at the sentence or discourse level. However, it remains to be seen if phonological elements can be combined in a single word, and how such phonological switching could be accounted for in a theory of bilingual phonology. We will present and discuss how ICS can be examined via experimental methodologies in order to provide the data needed to form the basis of any such account. Consistent with much other research on the mental lexicon, ICS reveals that lexical knowledge is a dynamic cognitive system which involves the interfaces of syntax, morphology, and phonology. We maintain that such knowledge and performance is well-modeled via the machinery of Distributed Morphology (DM). Furthermore, implementing a single word with multilingual components requires an understanding of how the linguistic systems interface with domains of general cognition, such as communication mode, executive control, and goal attainment. Moment-by-moment changes in the real-world environment influence the cognitive context of the speaker and explain the linguistic properties of the ICS speech. In this way, ICS data (perhaps more obviously than monolingual data) reveal how what used to be referred to as the "mental dictionary" is not a passive vocabulary repository to be retrieved but rather a networked, dynamic, distributed system. In our view, this is consistent with late-insertion, non-lexicalist models of morphology. DM offers a competence based, representational account of language which focuses on the well-formedness of grammatical structure (or *knowledge*). To address the production (or performance) side of things, we will adopt Sharwood Smith and Truscott's (2014) Modular Online Growth and Use of Language (MOGUL) model as it is primarily concerned with real-time performance and language *use*; the focus is on the production of grammatically acceptable utterances which transmit the desired meaning. By situating a competence-based model of grammatical structure (i.e., DM) inside a performance-based model of cognition (i.e., MOGUL), we hope to be able to account for how, what Chomsky calls I-language informs the E-language phenomenon of ICS.

In section *Review of Foundational Research on Intraword Codeswitching*, we provide an overview of the literature addressing the phenomenon of intraword codeswitching. In section *Discussion: Morphological Restrictions vs. Phonological Restrictions*, we look at the reported patterns of ICS, and probe the characteristics of both morphological and phonological switches. In section *Distributed Morphology*, we introduce the model of Distributed Morphology as the foundation for our accounts of the representational properties of ICS words and the generation of ICS words. Section *ICS at the Representational Level* summarizes the experimental techniques used to probe the question of whether switching phonology within a word is grammatical. Section *Producing ICS* introduces the MOGUL framework in order to model what underlies the *production* of an ICS. Section *Conclusion and Future Directions* provides our conclusions and future directions.

REVIEW OF FOUNDATIONAL RESEARCH ON INTRAWORD CODESWITCHING

In order to understand what type of phenomena we are attempting to explain, let us imagine another group of bilinguals. Fulana is a heritage Mexican Spanish speaker living in Chicago. She often gets together with friends and family who are all highly proficient in both Spanish and English. At times, their conversations are completely in English, at times completely in Spanish. Yet, there are also many instances where we can find a single sentence which contains both Spanish and English elements. You might hear Fulana say:

(a) *Siéntate Pedro*, you're going to spill your juice.
"Sit down, Pedro, you're going to spill your juice."

or

(b) Last week my *sobrina* came to visit.
"Last week my niece came to visit."

Clearly, these sentences reveal elements from both Fulana's Spanish and English. But what of the case of single words? Does Fulana also mix elements of her Spanish and English in a single word? The short answer is yes, Fulana might also say something, such as (c):

(c) Voy a *hang*ear con mis amigos
"I'm going to hang with my friends."

In this case, she combines the English verb "to hang" with Spanish verbal inflection to create a mixed or codeswitched word. Given Fulana's mixed use of Spanish and English, the question then becomes, how can we account for her use of elements from her two languages in the same way that we account for how a monolingual combines and uses elements from her one language. As this paper is concerned with CS at the word level, we limit our discussion to CS accounts that directly inform the use of two languages within a single word.

Of the foundational work which addresses ICS, most often cited are Poplack's (1980) *Free Morpheme Constraint* and MacSwan and Colina's (2014) *PF Interface Condition* (formerly realized as the *PF Disjunction Theorem*, MacSwan, 2000). Both the *Free Morpheme Constraint* and the *PF Interface Condition* claim that ICS is not possible. In her study of Spanish/English bilinguals living in NY, Poplack (1980) did not find many instances of mixed words. To explain their absence, she proposed the Free Morpheme Constraint which states "codes may be switched after any constituent in discourse provided that the constituent is not a bound morpheme" (pp. 585–586). Poplack claims that this strict constraint serves to account for the lack of occurrences of switches, such as (1), comprised of a root from L_A (here, English) and affixes from L_B (Spanish) in her corpus of Spanish/English CS.

(1) *eat-iendo

In (1), the English verb "eat" is combined with the Spanish bound affixes "-iendo." Following the *Free Morpheme Constraint*, a switch into another language cannot occur at this morpheme boundary and therefore the word in (1) is considered unacceptable to Spanish/English bilinguals and is not produced in bilingual discourse.

In a similar vein, the *PF Interface Condition* (2) also rules out intraword switches of the type shown above in (1).

(2) *PF Interface Condition*

i. Phonological input is mapped to the output in one step with no intermediate representations.
ii. Each set of internally ranked constraints is a constraint dominance hierarchy, and a language-particular phonology is a set of constraint dominance hierarchies.
iii. Bilinguals have a separately encapsulated phonological system for each language in their repertoire in order to avoid *ranking paradoxes*, which result from the availability of distinct constraint dominance hierarchies with conflicting priorities.
iv. Every syntactic head must be phonologically parsed at spell-Out. Therefore, the boundary between heads (words) represents the minimal opportunity for codeswitching.

In their formulation of the *PF Interface Condition,* which adopts a constraint-based (Optimality Theoretic, OT) perspective, MacSwan and Colina (2014) consider morphosyntactic X^0s (whether simple or complex) to be the input to PF/phonology. In order to avoid a ranking paradox of the phonological constraints between two languages, only a single phonology can be applied to a word (i.e., an X^0). In other words, PF does not allow for a word that has been formed in syntax to undergo a process in which the word is broken down into its individual morphological elements so that each element can undergo the phonological processes of its original phonological system and then be formed back into the original word. Instead, PF demands that a word formed in syntax will serve as the input to a single phonological system. Specifically, this input will yield a set of output candidates that will be evaluated by a (single) language-specific constraint ranking, thereby preventing phonological ICS.

It is essential to note that both Poplack (1980) and MacSwan and Colina (2014) recognize that codeswitched words in which one of the morphemes has been phonologically integrated into the other are attested in CS data. For instance, in (1), if "eat" is phonologically integrated into Spanish (i.e., with Spanish pronunciation [itiendo]), then it is considered acceptable to Spanish/English bilinguals. However, Poplack and MacSwan label instances of phonologically integrated mixed words as *borrowings* and claim that they arise as a result of a linguistic process distinct from CS. A borrowing (or "loanword"), can be defined as "a word that at some point in the history of a language entered its lexicon as a result of borrowing (or transfer, or copying)" (Haspelmath, 2009, p. 36). In a bilingual context, a borrowing is a word that has been taken from L_A and added to the mental lexicon of L_B, and, differently than CS, is typically morphologically, syntactically, and phonologically integrated into the recipient language (L_B). Because borrowings are accounted for differently than CS, Poplack and MacSwan posit that borrowings, such as [itiendo] do not serve as counter-evidence toward the *Free Morpheme Constraint* nor the *PF Interface Condition.* Now, before moving forward, we note that the purpose of this paper is not to comment upon the longstanding discussion of borrowing vs. CS nor to argue that any of the mixed words presented herein should be considered codeswitches instead of borrowings or vice-versa[1]. Instead, this paper serves as a critical discussion of the bilingual phenomenon of mixed words more generally and provides an overview of mixed words found in bilingual discourse. Furthermore, using phonological integration as the sole deciding factor for borrowings vs. CS (as is the case with [itiendo] and similar examples in the literature) is not optimal given that, at the mixed word level, borrowings can be phonologically indistinguishable from CS (i.e., they both demonstrate integration). For more discussion, see González-Vilbazo and López (2011), Poplack and Dion (2012), Bessett (2017), Grimstad (2017), Alexiadou and Lohndal (2018), among others.

In contrast to the *Free Morpheme Constraint* and *PF Interface Condition*, more recent work on ICS claims that intraword CS *is* possible but strictly constrained. For instance, Bandi-Rao and den Dikken's (2014) analysis of Telugu/English CS notes a difference in acceptability between the mixed word in (3) comprised of a Telugu root and English affixes and that of (4) comprised of a (mirror image) English root and Telugu affixes.

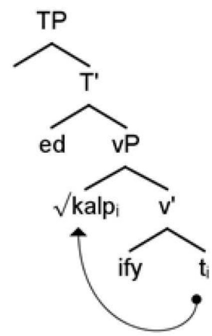
(3) my sister *kalp*-ified the curry
"my sister stirred the curry"

(4) *vaaDu nanni *love*-inc-EEDu
He-NOM me-ACC love-do-PST-AGR

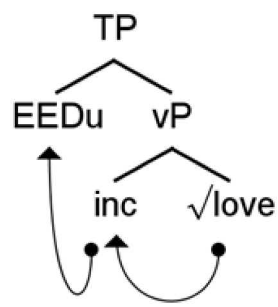
[1]As the mixed words presented in section *Discussion: Morphological Restrictions vs. Phonological Restrictions* are taken from an array of different sources which do not consistently provide the same information, we are, in fact, unable to comment upon this issue. In order to comment upon the borrowing vs. CS debate with respect to mixed words we argue that the very least the following four considerations should be taken into account: (1) morphosyntactic properties. For instance, does the root of the mixed verb behave morphosyntactically like a verb from Language A or B? (2) Phonological properties. What phonology(s) does the mixed word evince? (3) Demographic information on the bilingual who uttered the mixed word and the social and linguistic context in which it was uttered. (4) An analysis of the individual monolingual languages from which the mixed word was formed.

Bandi-Rao and Den Dikken claim that the difference between "kalpified" in (3) and "loveinceedu" in (4) is that "loveinceedu" was formed via incorporation in syntax and consists of a single morphosyntactic head, whereas "kalpified" in (4) was formed via a process of phrasal affixation and therefore is composed of two separate morphosyntactic heads. (5)–(6) illustrate this proposed difference in underlying structure[2].

(5)



(6)



To explain the difference in acceptability, Bandi-Rao and Den Dikken follow MacSwan (2000) and consider single morphosyntactic X^0s to be the input to phonology:

(7) Codeswitching within phonological words that are morphosyntactic heads (X^0s) is illicit.

Although (4) is argued to be illicit in a bilingual grammar because the mixed word consists of a single morphosyntactic head, the example in (3) suggests that ICS is possible as long as the underlying syntactic structure of the word in question is comprised of more than one morphosyntactic head. Each head in a mixed word, such as "kalpified" can receive its own (separate) phonology. However, because Bandi-Rao and den Dikken do not provide any phonetic information, we are unable to determine whether this word indeed demonstrates two phonologies.

González-Vilbazo and López (2011) also argue that ICS is possible but limited with respect to directionality and phonological form. Consider the German/Spanish mixed verbs in (8)–(9).

(8) *Utilis*-ier-en "we use"
use-v-3.PL.

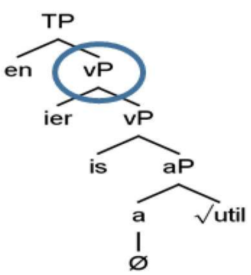
[2]Note that any and all syntactic trees presented in this paper are simplified representations for ease of explanation.

(9) **benutz*-ear "to use"
use-INF

(8) is a mixed word comprised of a Spanish verbal root/base *utilis* and German affixes *-ier-en*, whereas (9) is a mixed word comprised of a German verbal root *benutz* and Spanish affixes *-ear*. This directional asymmetry parallels that of (3)–(4) in Bandi-Rao and DenDikken in which only one directionality [Spanish to German in (8) and Telugu to English in (3)] gives rise to licit codeswitches while the opposite directionality results in ungrammaticality. Unlike Bandi-Rao and DenDikken, however, González-Vilbazo and López (2011), claim that switches, such as (9) are not possible due to a mismatch in features between the two languages[3]. They claim that Spanish little *v* has an unvalued feature for conjugation class[4] and that in the case of (9) it is unable to establish a syntactic dependency with the German root *benutz* because German roots do not have conjugation class features. In the case of (8), however, the conjugation class feature of the Spanish root is already valued and does not need to undergo any feature checking and is free to merge with the German affixes. Thus, the difference in features between Spanish and German verbs gives rise to the directionality asymmetry shown in (8)–(9)[5].

In addition to restrictions of directionality, González-Vilbazo and López (2011) also discuss restrictions on the phonological form of mixed words. They claim that "incorporation of a root into a suffix gives rise to an endocentric structure in which all and only the features of the head project to the newly created term" (p. 840). In the case of (8) let us take the German derivational affix *-ier-* to be the morphological head of the word as it is the highest derivational affix (10).

(10)



Following González-Vilbazo and López, it should be possible for the Spanish verbal base *-utilis* to incorporate into the German derivational affix *-ier*, giving rise to an output that is subject to the phonological rules of *-ier* (here, German). If this is the case, then the mixed word *utilisieren* is predicted to evince German phonology ([ʔʊtʰɪliːziːʁn] and is the attested output according to a Spanish/German bilingual consultant. Thus, based upon the observations made in Bandi-Rao and den Dikken (2014) and González-Vilbazo and López (2011) we see preliminary evidence that certain combinations of roots and affixes from different languages are possible. However, it might be the case that these morphologically mixed words, while having morphemes from two languages, evince a single phonology (that of the morphological head) instead of two phonologies. In order to determine if this is the case we need to take a closer look at the data (we return to this point in sections *Discussion: Morphological Restrictions vs. Phonological Restrictions* and *ICS at the Representational Level*).

Contrary to the accounts above, Jake et al. (2002) claim that phonologically mixed words *are* possible: Roots from L_A inflected with morphemes from L_B generally retain their L_A phonology (p. 75), as in (11)[6].

(11) Halafu m-tu-evaluate Swahili/English
then 2PL-1PL-evaluate
"Then you should evaluate us"
Note: *evaluate* is pronounced in English [ivæljueɪt], not as in Swahili [evaluete].

In (11) the English verb "evaluate" is merged with Swahili affixes. According to Jake et al., English phonology is maintained in the output of the inflected "evaluate," which suggests that (11) is a phonologically mixed word. With that said, MacSwan (2005) argues that *m-tu-evaluate* in (11) is formed by a process of phrasal affixation and is therefore composed of two underlying morphosyntactic X^0s instead of one [similar to the "kalpified" example provided by (Bandi-Rao and den Dikken, 2014) shown in (5)]. If it is the case that a single X^0 is the input to phonology, then *m-tu-evaluate* does not provide counterevidence toward the bans/constraints on ICS. We return to this issue in section *Discussion: Morphological Restrictions vs. Phonological Restrictions*.

In this section, we have provided an overview of the foundational work on intraword CS. While most (if not all) work on ICS agrees that there are certain restrictions on the ways in which morphological and phonological elements from different languages can be combined, a clear consensus as to what those restrictions are has not been reached (see also Alexiadou and Lohndal, 2018). In an attempt to clarify these restrictions, we have consolidated examples of word-internal CS from over 22 language pairs, which we present and discuss in the following section.

[3]González-Vilbazo and López (2011) also claim that mixed words, such as (8) are codeswitches while those, such as (9) are borrowings. In other words, in a bilingual context, Spanish/German bilinguals only utter mixed words comprised of Spanish roots and German affixes. Mixed words with German roots and Spanish affixes are only uttered in a Spanish monolingual context (between a Spanish/German bilingual and a Spanish monolingual with enough knowledge of German to understand the mixed word).

[4]The three conjugation classes in Spanish are: AR, ER, IR.

[5]We note that this feature checking analysis is unable to explain the directionality asymmetry in Spanish/English CS. In Spanish/English CS, English roots can combine with Spanish affixes, but Spanish roots cannot combine with English affixes (i.e. the opposite order from Spanish/German). If it is the case that English roots, like German, do not have conjugation class features, then how does the Spanish v become valued in the case of words like *dipear* "to dip," *mopear* "to mop," *parquear* "to park"?

[6]See Jake et al. (2002) for more examples and MacSwan (2005) for critique and discussion of whether the examples serve as counterevidence toward the PF Disjunction Theorem.

DISCUSSION: MORPHOLOGICAL RESTRICTIONS VS. PHONOLOGICAL RESTRICTIONS

In this section, we systematically explore patterns and trends found among 57 examples of ICS (**Table 1**) to better understand the nature of their morphological and phonological restrictions. In other words, we attempt to explain what is meant when researchers say "intraword CS is *not possible*" or "intraword CS is *sharply limited*" based on these data. We also refer the reader to Alexiadou and Lohndal (2018) for a similar discussion of morphological restrictions on ICS. In **Table 1**, the root of each mixed word is italicized, and any affixes are separated by dashes. Ungrammatical or unacceptable mixed words are denoted by an asterisk.

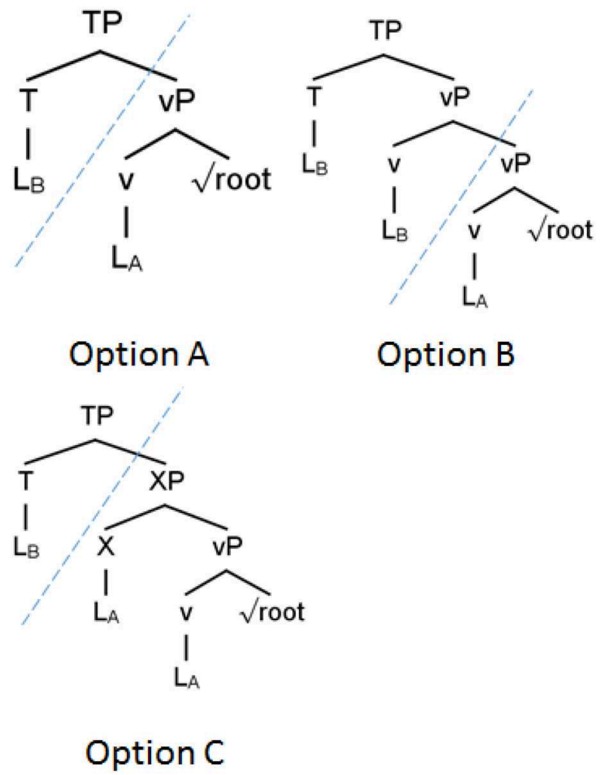
Before moving forward, it is important to note that almost all of the data in **Table 1** come from corpora. Data from corpora allow us to examine what bilinguals produce but do not allow us to determine what is illicit or impossible in a bilingual grammar. Just because a bilingual does not *produce* a certain construction, does not mean that it is not possible for that construction to be produced in a different context in a way that would be deemed acceptable by speakers from the relevant community. Studies employing a methodology that targets negative evidence, or what is *not* possible, are best equipped to answer these questions that arise from analysis of corpus data. As an example, consider an acceptability judgment task or a forced-choice task which exposes bilinguals to mixed words with switches between (a) categorized roots and derivational affixes, (b) different types of derivational affixes, and (c) derivational affixes and inflection. Judgments of these switch types could provide experimental evidence for the morphological restrictions on word-internal CS. These types of studies are often done in conjugation with a syntactic analysis of CS across word boundaries (e.g., Bartlett and González-Vilbazo, 2013; González-Vilbazo and Koronkiewicz, 2016) and should be extended to ICS as well.

As our review of ICS accounts presented in section *Review of Foundational Research on Intraword Codeswitching* suggest that phonological restrictions on ICS exist separately from morphological restrictions, we treat the morphology and phonology of ICS as two (potentially) unrelated/separate phenomena. In other words, we do not assume that a morphologically mixed word necessarily precludes a phonologically mixed word (see Stefanich, 2019 for further discussion). We begin with morphological aspects, followed by phonological aspects.

Trends in the Morphology of ICS

In **Table 1** we see 39 examples of morphologically mixed verbs, 16 examples of morphologically mixed nouns, and two examples of mixed adjectives. Upon first glance, we can see several surface patterns that allow us to explore the following questions: (1) between which morphemes do switches occur? and (2) in which direction do switches occur? Following a theory of DM (see section Distributed Morphology), we can posit that we might see switches occur at the following morpheme boundaries: (a) categorized root + inflection, (b) categorized root + categorizing head/derivational affix, (c) inflection + inflection, and in the following directions: (a) L_A to L_B and (b) L_B to L_A. These different options are illustrated in (12).

(12)



Between Which Morphemes Do Switches Occur?

In general, we see two surface patterns that arise from the data in **Table 1**[7]:

A) lexical roots tend to come from one language (L_A) while affixes, both derivational and inflectional, come from another (L_B).

More specifically, in the case of the verbs, we see a switch boundary that occurs between the lexical root and a derivational affix (e.g., little *v*) or some sort of verbalizing affix [see examples (1)–(2), (11)–(12), (19), (21), (34)–(36) in the table]. The derivational/verbalizing affix tends to be productive in the language of origin (i.e., the affix used to make new verbs in that language). In (1)–(2) this affix is German *v -ier*, used to verbalize Latinate roots. In (3)–(6) it is Spanish "-ear" and in (25)–(26) it is Dutch *-er*. In (11)–(12) and (19)–(21) it is Hungarian *-ol*, and in (34)–(36) Greek *-ar*.

B) Switches are not attested between the affixes themselves (i.e., between derivational and inflectional affixes or between inflectional affixes).

Instead, the affixes come from a single language, which aligns with the claim by González-Vilbazo (2005) and López et al.

[7]We note that the surface patterns presented here align with those discussed in Alexiadou and Lohndal (2018).

TABLE 1 | A crosslinguistic survey of intraword codeswitching.

Source	Language pair	Example			Morphology	Phonology
VERBS						
González-Vilbazo and López (2011)	German/Spanish	(1) *utilis*-ier-en use-v-1PL "we use"	(2) *alemanis*-ier-t germanize-v-3SG "it germanizes"		Spanish root + German affixes	German
González-Vilbazo and López (2011)	German/Spanish	(3) **benutz*-ear use-INF "to use"	(4) **lauf*-ear run-INF "to run/walk"		German root + Spanish affixes	Spanish
MacSwan (2005)	Spanish/English	(5) *it*-eando eat-PROG "eating"	(6) *it*-ar-á Eat-FUT-3SG "will eat"		English root + Spanish affixes	Spanish
Stefanich and Cabrelli Amaro (2018a)	Hindi/English	(7) *dhadk*-oing beat-PROG "beating"			Hindi root + English affixes	English
Akinremi (2017)	Igbo/English	(8) *wed*-ì-rì wed-EV-past "wedded"	(9) *work*-ù-ghì work-EV-NEG "not work"		English root + Igbo affixes	Igbo
Hlavač (1999)	Croatian/English	(10) *pak*-ujem pack-1SG.PRS "I pack"			English root + Croatian affixes	Croatian
Bolonyai (2005)	Hungarian/English	(11) fel-*réz*-ol-t-am PV/up-raise-VBZ-PST-1SG "raised up"	(12) *fájndaut*-ol-j-a ki find out-VBZ-IMP-3S PV/out "find it out"		English root + Hungarian affixes	Hungarian
MacSwan (2005)	Spanish/English	(13) **eat*-iendo eat-PROG "eating"	(14) **eat*-ar-á eat-FUT-3SG "will eat"		English root + Spanish affixes	Mixed
Alexiadou (2017)	Greek/German	(15) **Kampf*-ar-o fight-AFF-1SG "I am fighting"	(16) **schwim*-ar-o swim-AFF-1SG "I am swimming"		German root + Greek affixes	Mixed
Jake et al. (2002)	Swahili/English	(17) m-tu-*evaluate* 2P-1PL-evaluate "evaluate"	(18) si-ku-*come* 1SG.NEG-PST.NEG-come "I didn't come"		English roots + Swahili affixes	Mixed
Bolonyai (2005)	Hungarian/English	(19) *order*-ol-t order-VBZ-3SG.PST.INDEF "ordered"	(20) **lunch*-t-unk lunch-PST- 1PL.INDEF "lunched"	(21) *lunch*-ol-t-unk lunch-VBZ-PST-1PL.INDEF "lunched"	English root + Hungarian affixes	Mixed
Bandi-Rao and den Dikken (2014)	Telugu/English	(22) *kal(i)p*-ifi-ed stir-v-PST "stirred"			Telugu root + English affixes	Not reported (mixed)
Bandi-Rao and den Dikken (2014)	Telugu/English	(23) **love*-inc-eedu love-do-PST.AGR "loved"			English root + Telugu affixes	Not reported (mixed)
MacSwan (2000)	Nahuatl/Spanish	(24) nik-*amar*-oa 1S.3Os-love-VSF "love"			Spanish root + Nahuatl affixes	Unclear (Nahuatl)
Treffers-Daller (1993)	Dutch/French	(25) *offr*-er-en offer-v-INF "to offer"	(26) *traduis*-er-en translate-v-INF "to translate"		French root + Dutch affixes	Unclear (Dutch)
Fuller (1999)	German/English	(27) ge-*farm*-t PTCP-farm-PST "farmed"	(28) ge-*move*-t PTCP-move-PST "moved"		English root + German affixes	Unclear (German)
Halmari (1997)	Finnish/English	(29) *pretend*-at-tiin pretend-V-PASS-PST. "pretended"			English root + Finnish affixes	Unclear (Finnish)

(Continued)

TABLE 1 | Continued

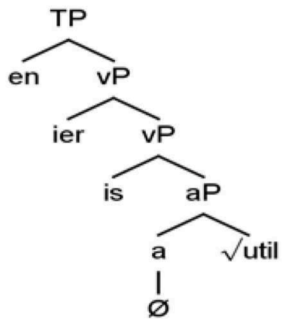
Source	Language pair	Example	Morphology	Phonology
Grimstad et al. (2014)	Norwegian/English	(30) *teach*-a teach-PST “taught” (31) *spend*-a spend-INF “to spend” (32) *rent*-er rent-3SG “rents” (33) *walk*-te walk-PST “walked”	English root + Norwegian affixes	Not reported
Alexiadou (2017)	Greek/English	(34) *muv*-ar-o move-AFF-1SG “I am moving” (35) *kansel*-ar-o cancel-AFF-1SG “I am canceling”	English root + Greek affixes	Not reported
Alexiadou (2017)	Greek/German	(36) *skan*-ar-o scan-AFF-1SG “I am scanning”	German root + Greek affixes	Not reported
Bokamba (1989)	Lingala/French	(37) a-*téléphon*-aka AGR-call-PRS “calls”	French root + Lingala affixes	Not reported
Stammers and Deuchar (2012)	Welsh/English	(38) *exfoliate*-io exfoliate-NONFIN “exfoliate” (39) *emphasize*-io emphasize-NONFIN “emphasize”	English root + Welsh affixes	Not reported
NOUNS				
Hlavac (2000)	Croatian/English	(40) *kontejner*-e container-M.PL.ACC “containers”	English root + Croatian affixes	Croatian
Halmari (1997)	Finnish/English	(41) *napkin*-eita napkin-PL “napkins” (42) *workshopp*-i-a workshop-SF-PART “workshop”	English root + Finnish affixes	Unclear (mixed)
Grimstad et al. (2014)	Norwegian/English	(43) *grad(e)*-en grade-DEF.M.SG “grade” (44) *birthday*-en birthday-DEF.M.SG “birthday”	English root + Norwegian affixes	Not reported
Turker (2000) in Jake et al. (2002)	Turkish/Norwegian	(45) *sentrum*-a center-DAT “center” (46) *forelesning*-ler-e lecture-PL-DAT “lectures”	Norwegian root + Turkish affixes	Not reported
Backus (1992) in Jake et al. (2002)	Turkish/Dutch	(47) *meisje*-den girl-ABL “girl”	Dutch root + Turkish affixes	Not reported
Amuzu (1998) in Jake et al. (2002)	Ewe/English	(48) *topic*-a e-wo topic-PL-INDEF “some topic”	English root + Ewe affixes	Not reported
Cantone (2003)	Italian/German	(49) *topf*-ino pot-DIM “little pot”	German root + Italian affixes	Unclear
Cantone (2003)	Italian/German	(50) *cas*-en house-PL “houses” (51) *gelat*-en ice-cream-PL “ice-creams”	Italian root + German affixes	Unclear
Riksem et al. (2019)	Norwegian/English	(52) *mower*-e mower-INDEF.PL.M “mowers” (53) *farmer*-ne farmer-DEF.PL.M “the farmers”	English root + Norwegian affixes	Not reported
Budzhak-Jones (1998)	Ukranian/English	(54) *atment*-iv atments-GEN “atments” (55) *stor*-a store-M.GEN “(the) store”	English root + Ukranian affixes	Not reported
ADJECTIVES				
Treffers-Daller (1993)	Dutch/French	(56) *violent*-e violent-F “violent”	French root + Dutch affixes	Dutch
Treffers-Daller (1993)	Dutch/French	(57) *sympathiqu*-e nice-F “nice”	French root + Dutch affixes	Mixed

The root of each mixed word is italicized, and any affixes are separated via dashes. Ungrammatical or unacceptable mixed words are denoted with an asterisk. Not reported, no mention of phonology; Unclear, phonology is discussed in some aspect related to the examples but we are unable to determine based upon what is mentioned.

(2017) that morphological switches between derivational and inflectional affixes are not possible.

However, surface patterns and generalizations are often misleading. We discuss three examples, (1) and (52)–(53), that challenge the generalizations presented above. First, in (8) [from (1) in the **Table 1**], the Spanish/German mixed verb *utilisieren*, let us consider the proposed underlying structure in (13) (cf. Alexiadou and Lohndal, 2018 for an alternative analysis).

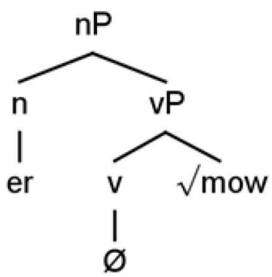
(13)



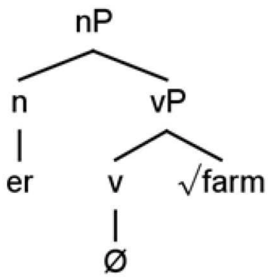
The Spanish part of this mixed word, *utilis*, which comes from the Spanish verb *utilizar*, is comprised of an adjectival root *util* that is merged with a Spanish little *v* (spelled out here as -*is*). If this analysis is correct, then the mixed word *utilisieren* demonstrates a morphological switch between two derivational affixes, here Spanish *v* and German *v*. While this example does not directly contradict the generalizations in A-B, it does point out two important things. The first is that for generalization A it would be erroneous to assume that whatever comes from L_A is solely the root. The underlying structure of the morphological elements that come from L_A could contain an already complex structure. Second, this example illustrates that it is not always the case that all of the affixes come from the same language. If the underlying structure proposed in (13) is correct, then it looks like a derivational affix from one language can be merged with a derivational affix from another.

Now, let us consider (52)–(53) in which we have the Norwegian/English mixed nouns *mowere* "mowers" and *farmerne* "the farmers." On the surface it appears that the English nouns "mower" and "farmer" are merged with Norwegian inflection. However, just like with *utilisieren*, we can break down the components of the elements from L_A, here English. Consider that the English nouns "mower" and "farmer" have the underlying structure shown in (14).

(14a)



(14b)



That is, the English nouns "mower" and "farmer" are complex nouns; the verbs "mow" and "farm"[8] are merged with the English derivational affix "er" to make them nouns[9]. If we assume the structure in (14), then the mixed words *mowere* and *farmerne* demonstrate morphological switches between a derivational affix (here English *n*) and Norwegian nominal inflection. Such a switch boundary contradicts generalization B, which precisely states that there should be no switching between derivational and inflectional affixes. The examples in (52)–(53) further demonstrate that we cannot just rely on surface level patterns, but that we must take a look at the underlying structure of the elements from both L_A and L_B.

For instance, Alexiadou and Lohndal (2018) discuss the Spanish/German mixed word *segurat*-en "security men," comprised of a Spanish root and a German plural affix, in their cross-linguistic analysis of ICS in the nominal domain. They point out that, in contrast to mixed words in the verbal domain, "*segurat*en" does not have overt nominalizing morphology. Alexiadou and Lohndal suggest that there is a covert nominalizing affix (i.e., *n*) that categorizes the Spanish root and makes it a noun, which is then able to merge with a German plural affix. Following our current line of discussion, the pertinent question here is whether this covert nominalizing affix is a Spanish *n* or a German *n*[10]. This difference in underlying structure is important, because one option (i.e., German *n*) falls in line with the generalization that there can be no switching between affixes, while the other (i.e., Spanish *n*) contradicts it.

Further, consider an additional question born out of examples (1), (52)–(53): if switching between derivational affixes is possible [as seems to be the case in (1)], are there any restrictions on the type or direction of the derivational affixes? For instance, in (1) the switch boundary occurred between Spanish *v* and German *v* (i.e., derivational affixes of the same type). Would it be possible to switch between, say, a Spanish *a* and German *v*? Or between Spanish *n* and German *v*? In (52)–(53) we see a similar pattern of symmetry between affixes, an English *n* merged with nominalizing Norwegian inflection. That is, even though we see a switch between a derivational affix and inflection, the inflection used is of the type required by the category of the word (here, noun). In order to answer these questions as to the licit morphological switch boundaries, experimental methodologies that directly examine ICS with different underlying structures are needed.

[8]We note that the word "farm" in English lacking any context could be both a noun and a verb, and that it is potentially ambiguous whether the underlying structure of the verb "to farm" includes the noun or not. For the purposes of our paper whether "farm" maintains the more complex structure or not does not affect the morphological switch point under discussion. See Acquaviva (2009), Alexiadou and Lohndal (2017), Borer (2013), and Embick (2015), among others for a discussion on the categorization of roots.

[9]The derivational affix "er" is a productive affix in English that is used to convert verbs to nouns with the meaning "one who [verb]s."

[10]We use "Spanish" and "German" here as descriptive labels for the underlying features that comprise a nominalizing affix in the Spanish language versus the German language.

In Which Direction Do Switches Occur?

Another pattern evident within the data in **Table 1** is that of a directional asymmetry in how the morphemes are combined. This asymmetry, which has also recently been reviewed in Alexiadou and Lohndal (2018), supports the examples discussed in section *Review of Foundational Research on Intraword Codeswitching*; it seems to be the case that mixed words can be composed of roots from L_A and affixes from L_B but that the reverse is not possible. This asymmetry is attested/claimed for language pairs, such as Spanish/German (González-Vilbazo and López, 2011), Telugu/English (Bandi-Rao and den Dikken, 2014), Greek/English and Greek/German (Alexiadou, 2017), French/Dutch (Treffers-Daller, 1993), Spanish/English (Stefanich and Cabrelli Amaro, 2018a), and Norwegian/English (Grimstad et al., 2014; Riksem et al., 2019).

Of the 57 examples collected in **Table 1**, there is one instance in which we see possible counter-evidence toward this directionality asymmetry. The Italian/German mixed nouns in (50)–(51) comprise Italian roots and German affixes, while the noun in (49) comprises a German root and Italian affixes. The reported acceptability of these mixed words suggests that a directionality asymmetry in word-internal CS is not universal, but rather that it most likely depends on the feature combinations of the language pair itself. However, these examples come from children between 3 and 4 years of age, in which it could be the case that these bilinguals are still in the process of acquiring the relevant German and Italian features. Moreover, the children were reported to favor switches when in Italian mode over German mode. Thus, the attested bidirectionality of their switches may not be part of an adult bilingual grammar and should be confirmed with adults via methodologies, such as those discussed previously (acceptability judgment task, forced choice, etc.) to tease apart any confounding factors related to language acquisition.

Different analyses have been proposed to account for this directionality asymmetry in ICS (recall discussion of González-Vilbazo and López, 2011 and Bandi-Rao and den Dikken, 2014 in section *Review of Foundational Research on Intraword Codeswitching*), with some scholars maintaining that asymmetry is a characteristic of CS more generally (e.g., Myers-Scotton, 1992 et seq). In their review of this observed directional asymmetry, Alexiadou and Lohndal (2018) point out that speakers tend to use the default overt realizations of verbalizing affixes (e.g., *ar* in Greek, *ier* in German), and they suggest that this tendency might result in this asymmetry. In other words, when codeswitching with a language pair where L_A has default overt realizations of v and n but L_B does not, the affixes will come from L_A. However, this account would not be able to explain what happens in language pairs where either both languages or neither language demonstrates default overt realizations. As we suggest in section *Discussion: Morphological Restrictions vs. Phonological Restrictions*, making use of experimental methodologies beyond corpora analysis will put us on the path toward answering some of these questions raised by evaluation of these corpora.

Experimental methodology aside, the examples in **Table 1** demonstrate that intraword morphological switches are possible but constrained/limited in a systematic way. This is representative of CS as whole, which is considered a systematic and rule-governed phenomenon, the same as any monolingual grammar. We now turn to a discussion of the phonological aspects of ICS.

Trends in Phonology of Word-Internal CS

Phonological switches seem to behave differently than morphological switches at the word level. As previously noted, Bandi-Rao and den Dikken (2014), MacSwan (2000), and MacSwan and Colina (2014) claim that ICS is not possible due to requirements of phonology and thus phonological outputs are predicted to not contain phonological elements from two languages. Looking at our 57 examples in **Table 1**, is their prediction confirmed, or are there examples of phonologically codeswitched words? Unfortunately, it is not so easy to answer that question. First, phonology is not addressed in most studies on ICS, which makes it either impossible to determine what the phonology of the mixed word is, or at best we must infer from authors' indirect remarks. When phonology *is* addressed, it is done so in an anecdotal manner or based solely on impressionistic analysis and lacks acoustic information or experimental data.

Second, any phonological analysis provided for the mixed words in **Table 1** lacks the bilingual source's monolingual productions as a point of comparison. Just as a growing body of research on syntax and CS uses a bilingual's own monolingual judgments as a control measure for CS data to account for individual variation and language contact (see González-Vilbazo et al., 2013; Ebert and Koronkiewicz, 2018 for a discussion) so must research on phonology and CS.

Third, we must clearly define the parameters used to define the term "word." Recall from section Review of Foundational Research on Intraword Codeswitching that the accounts that suggest a ban on phonological ICS claim that the restriction only applies to phonological words that are comprised of single morphosyntactic heads (X^0s, e.g., verbs whose affixes incorporate into the root). It could be the case that phonological ICS is permitted in phonological words that are comprised of two separate X^0s. Thus, when analyzing the phonology of mixed words, it is important to also look at them with respect to their underlying syntactic structure in order to identify and establish a more refined view of the restriction on phonological switches. With that being said, what can we glean from the data in **Table 1**? Twenty-three examples in **Table 1** were provided by authors with a phonological description, 13 of which are said to demonstrate a single phonology. Of the remaining 34 examples, 27 do not contain any mention of phonology. While the other seven do not provide any explicit phonological description, we can make an educated guess based upon the authors' discussion. We discuss each set in turn.

Single Phonology

According to their sources, examples (1)–(12), (40), and (56) all demonstrate a single phonology; in each case, the phonology of the mixed word matches that of the language of the affixes and not the language of the root. For example, in (1), the affixes are German and the phonology is German, as represented by sounds, such as a glottal stop, a high back rounded vowel and a voiced alveolar fricative. In (8)–(9), the affixes are Igbo, and the phonology is Igbo, as represented by vowel harmony, tone and stress. In (56), the affix is Dutch and the phonology is Dutch in that the French nasal vowel has been replaced by a Dutch vowel. This observation that the phonology of a mixed word will come from the language of the affixes is essential because it makes testable predictions for experimental research on ICS (see section ICS at the Representational Level for an overview of one such study). Further, note that this observation falls in line with the work of González-Vilbazo and López (2011) presented in section *Review of Foundational Research on Intraword Codeswitching*. González-Vilbazo and López claim that the morphological head of the word projects its features to the whole word. It is likely that the morphological head of these mixed words is the highest derivational affix, so it logically follows that the phonology of the mixed words matches that of the affixes (see Stefanich, 2019 for further discussion).

Mixed (Two) Phonologies

Contrary to the examples of mixed words that demonstrate a single phonology, there are ten examples (13)–(21), (57) in **Table 1** whose sources state that they are instances of mixed phonological words. We can divide these examples into two groups: (1) considered unacceptable or not licit in a bilingual grammar and (2) considered acceptable or licit in a bilingual grammar. First, (13)–(16) are morphologically mixed words comprised of English/German roots and Spanish/Greek affixes. The roots and affixes each maintain their "donor" language phonology (e.g., the English root has English phonology, but Spanish affixes have Spanish phonology). These phonologically mixed words are considered unacceptable/ungrammatical according to the authors, lending support to the constraints presented in section *Review of Foundational Research on Intraword Codeswitching* that ban phonologically mixed words.

In contrast to (13)–(16) the other six examples (17)–(21), (57) are considered acceptable/grammatical by the authors and thus could constitute possible counterevidence toward the constraints on phonological ICS discussed in section Review of Foundational Research on Intraword Codeswitching. The mixed word in (17) is comprised of the English verb "evaluate" and Swahili inflection. Jake et al. (2002) claim that the English verb retains its English phonotactics instead of demonstrating Swahili phonotactic (CVCV) and nucleus structure: The final syllable is closed and the front vowels have off-glides. In (57), we see a French adjective *sympathique* "nice" combined with a Dutch agreement affix-*e*. Treffers-Daller (1993) claims that while the French adjective is pronounced as it would be in monolingual French, citing the presence of a nasal vowel as evidence, the Dutch affix is pronounced as a schwa, as it would be in Dutch[11]. The last examples (19)–(21) demonstrate mixed words with English roots and Hungarian affixes. Bolonyai (2005) claims that these mixed words are grammatical and that the English root maintains its English phonology while the Hungarian affixes demonstrate Hungarian phonology. However, she does not provide transcription or acoustic detail.

Pending that acoustic evidence would back up the authors claims that the examples (13)–(21) and (57) are words that demonstrate two phonologies can we say then, that they constitute counterevidence toward the ban on phonological intraword CS addressed in section *Review of Foundational Research on Intraword Codeswitching?* MacSwan (2005) addresses examples (17) and (57) and claims that *mtuevaluate* and *sympathique* are actually two separate morphosyntactic X^0s formed by a process akin to phrasal affixation and not a single morphosyntactic X^0 [similar to (5) in section *Review of Foundational Research on Intraword Codeswitching*]. As such, these words are "allowed" to have two phonologies (assuming each X^0 can demonstrate its own phonology)[12]. Thus, following MacSwan's analysis, examples (17) and (57) would not constitute counterevidence toward the ban on phonological intraword CS discussed in section *Review of Foundational Research on Intraword Codeswitching*.

Following this line of thought, we ask whether examples (19)–(21) are (a) mixed words formed by a process of incorporation [like (1)] where the output is a single but complex X^0 that is sent to phonology as one unit (and therefore can only evince one phonology), or (b) formed by a different process in which the output is two separate X^0s, each of which can be sent to a different phonology, therefore giving rise to what appears to be a phonologically mixed word. Bolonyai (2005) claims that English verbs are integrated into Hungarian via a "derivational, denominal verbalizer suffix" (p. 317). In (19) and (21) this suffix is realized overtly as *ol* and (20) demonstrates that without this overt suffix the combination of an English root with Hungarian inflection results in ungrammaticality. This analysis suggests that the mixed verbs in (19) and (21) are formed via a process of incorporation and therefore should be subject to the constraint on intraword phonological switching. A more in-depth analysis of verb formation in Hungarian is needed to confirm.

Additionally, note that (19)–(21) stand in contrast to the English/Hungarian examples (11)–(12). While the mixed words in (11)–(12) demonstrate Hungarian phonology, the words in (19)–(21) reportedly demonstrate mixed phonology (here, English in the root and Hungarian in the affixes). We then ask, why would it be the case that sometimes the mixed words demonstrate Hungarian phonology and sometimes a combination of English and Hungarian phonologies?[13] Further, Bolonyai states that (19)–(21) are "morphologically integrated

[11]We recognize that there are some varieties of French that may produce word-final schwa. If this word is uttered by a bilingual from one of those varieties, then the production of the Dutch affix as a schwa would not be evidence of a phonological switch.

[12]We acknowledge that MacSwan (2005) does not provide a detailed syntactic analysis of these examples and that in order to say definitively that they are two separate X0s, such an analysis is required.

[13]Bolonyai (2005) appeals to a borrowing versus codeswitching account to explain the difference in phonological realization between examples, such as (11)–(12) and (19)–(21). However, she does not provide any details regarding the bilingual speakers who uttered these mixed words nor any information as to the context in which they were uttered. Without this information it is difficult to address whether a borrowing versus codeswitching account holds.

forms that occur with no (or minimal) phonological assimilation to Hungarian (i.e., there appears to be conscious retention or approximation of English pronunciation)" (p. 318). What is meant by "conscious retention" here? If phonologically mixed words are licit in a bilingual grammar, then a bilingual should not be conscious of the fact that s/he is "retaining" a specific phonology; if s/he is, such a production is reflective of metalinguistic knowledge rather than his/her bilingual grammar.

Unclear or Not Reported Phonology

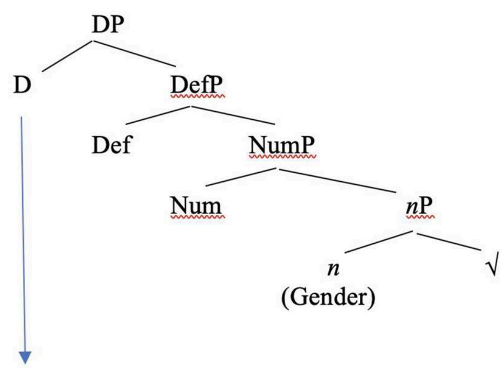
For the remainder of the examples in **Table 1**, the phonology is either unclear or not explicitly addressed. For instance, in (27)–(28) the phonology is not addressed directly, but the author points out that these words demonstrate final obstruent devoicing, a phonological process that occurs in German (the language of the affixes) and not English (the language of the root). This suggests that it might be the case that (27)–(28) demonstrate German phonology; however, without explicit mention of the phonology of the root we are unable to make that claim. In a few cases, we are able to make an educated guess based upon the discussion of the examples in the original source. For example, we can infer that morphologically mixed words in (22)–(23) are also phonologically mixed words based upon the constraint Bandi-Rao and den Dikken (2014) posit to account for these words (see section *Review of Foundational Research on Intraword Codeswitching*). Further, if the mixed words are labeled as borrowings and the author(s) assumes a traditional view of borrowing in which borrowings demonstrate phonological integration, then we can assume that the word in question evinces a single phonology [e.g., (24)].

As seen in this section, we are able to see (surface) trends in the morphology of mixed words. We must remember that surface phenomena do not always reveal the nature of the underlying representations in grammar when it comes to such things as morpheme order or word order. This is true of CS as well. However, there remains a (semi) open debate as to whether we have concrete examples of a single word evincing two phonologies. As stated, any attempts to examine the phonology of a mixed word must first set a clear definition for what is considered a "word" and whether any constraints apply strictly to morphological (one X^0) vs. phonological words (one or more X^0s). Second, an acoustic analysis of the phonology of the mixed word, as well as of the two "monolingual" phonologies involved is essential. Nevertheless, our examination of the examples in **Table 1** did reveal (minimally) 2 distinct patterns. On the one hand, mixed words composed of a single X^0 seem to demonstrate a single phonology, namely that of the affixes. On the other hand, mixed words that are composed of two separate X^0s seem to possibly demonstrate two phonologies. In the next section we present a candidate for a theoretical account of intraword CS.

DISTRIBUTED MORPHOLOGY

The central concept of DM (Marantz, 1997; Arad, 2003; Embick and Noyer, 2007; Lohndal, 2013; Grimstad et al., 2014; Harley, 2014) is that a single generative engine governs sound/meaning correspondences, making no distinction between word-level and phrase-level syntax. As McGinnis (2016) notes, "DM departs from the traditional notion of the Saussurean sign, which directly associates a phonological form with a meaning. Instead, the theory postulates that the stored knowledge of a language is distributed across three separate lists" (p. 390). One list is known as the Lexicon. The Lexicon is where elements which are found on the terminal nodes of a syntactic tree are stored. These elements can be either lexical roots or grammatical morphemes. Both inflectional and derivational morphemes are made up of bundles of syntactic/semantic features. However, a content morpheme is represented by a category-neutral lexical root. At this stage of a derivation, there is no phonological content to the morphemes. Phonological content is added later by reference to the second list: the Vocabulary. Vocabulary items are inserted into terminal nodes of the syntactic derivation after Spell-out (hence *late* insertion). The third list is the Encyclopedia, which associates lexical roots post-syntactically with special, non-compositional aspects of meaning. Of particular relevance here, DM offers a model where under-specified morphological elements compete for late-insertion into a fully generated syntactic tree, complete with syntactic terminal nodes. Let us consider an example which can show how a Determiner might spell out varying morphosyntactic features. A Spanish DP includes features for definiteness, number and gender, as shown in the tree in (15):

(15)

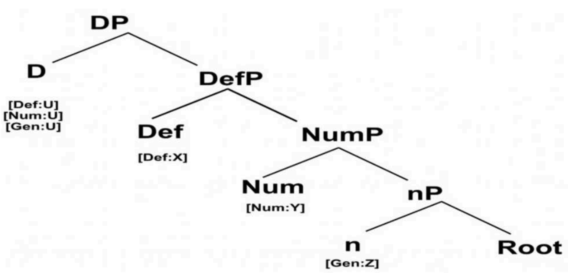


F. PL. DF → las
PL. DF → los
F. DF → la
M. DF → el
DF → lo

These different Determiners are in competition for insertion into the terminal node. The vocabulary item which matches the most features without being overspecified (i.e., having more features than necessary) will be inserted into the tree.

Now, as we will be looking in depth at the performance of Norwegian/English CS in section Producing ICS, we are going to use Norwegian to demonstrate how DM works. Following Grimstad et al. (2014; page 224), a Norwegian DP would be as shown in (16):

(16)



The syntactic terminal list contains two types of primitives: category neutral lexical roots (i.e., √TABLE, √CAT, √RUN, etc.) and grammatical morphemes. While, the exact nature of roots in DM is still a topic of debate (Harley, 2014), this paper will follow Grimstad et al. (2014) in assuming that roots contain no grammatical features themselves and are underspecified both phonologically and semantically.

The tree in (16) has specified syntactic features but no phonological or semantic content. The phonological items of a word which match the abstract features in the template can then be inserted into the derivation. This process is known as *vocabulary insertion*. For Norwegian, the possible vocabulary items would include those in (17):

(17) M.SG.DF -> -en
F.SG.DF -> -a
N.SG.DF -> -et
PL.DF -> -ene
PL. -> -s

Schematically, we could present the stages in a derivation as given in **Figure 1**.

Much work on DM has been concerned with the spell-out of functional morphemes (though see Archibald, 2016 and (Haugen and Siddiqui, 2013), for a discussion of competition for roots). Here Vocabulary items compete for insertion, subject to the Subset Principle (Halle, 2000).

> **Subset Principle** (Halle): The phonological exponent of a Vocabulary Item is inserted into a position if the item matches all or a subset of the features specified in that position. Insertion does not take place if the Vocabulary Item contains features not present in the morpheme. Where several Vocabulary Items meet the conditions of insertion, the item matching the greatest number of features specified in the terminal morpheme must be chosen.

Two principles are key:

A. Only Vocabulary Items which specify a subset of a head's features can be inserted
B. Only the most specific Vocabulary Item is inserted

In English (De Belder and Van Craenenbroeck, 2015), a DP would consist of the D-Head features and a category-neutral Root, as shown in (18).

(18) DP [[+D,+def]√]

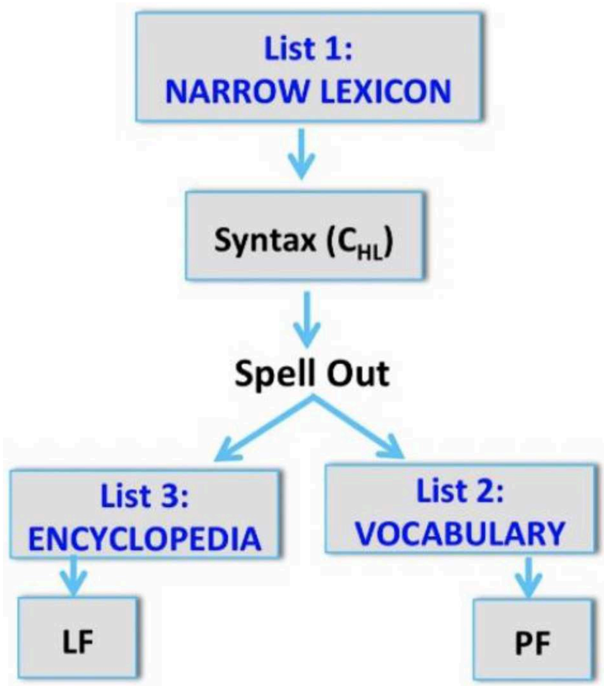


FIGURE 1 | Mapping sound onto meaning (Burkholder et al., 2017).

When the syntactic derivation is completed, this structure is handed over to the post-syntactic component responsible for pronouncing it. Now Vocabulary Insertion takes place: the terminal nodes in the syntactic structure in (18) need to be matched up with appropriate lexical exponents. This means that the post-syntactic Vocabulary will contain correspondence rules, such as the ones in (19).

(19) [+D,+def] <-> /ðə/ b. √ <-> /bʊk/

Given that the features on the left in the Vocabulary List match the features in the terminal node of the syntactic tree perfectly, then [ðə] can be inserted.

Let's consider an example from German to probe more issues of competition. In German it is reasonable to assume that there is one nominal plural suffix in the syntax; hence plural nouns may all have the abstract syntactic representation: [[NOUN]-PL]. However, the German vocabulary provides a variety of vocabulary items that express this node, including: -*Ø, -(e)n, -e, -er* and-*s* (ignoring some changes that might occur in the stem). These allomorphs would be in competition for insertion into the tree. The competition would be governed by the subset principle, by stem-conditioned associations, and by familiar morphological blocking conditions which ensure that the insertion of a more-specified vocabulary item blocks the insertion of a less-specified one.

McGinnis (2016) also illustrates matters of allophonic competition. For example, in DM, the alternating forms of the English plural shown in (20a–c) are taken to realize the same syntactic nodes—minimally, a lexical root and a node bearing number features, which has several possible morphological realizations. (20a) shows the unrestricted default plural allomorph, which is also subject to phonologically conditioned variation (as in cat-[s], dog-[z], horse-[əz]). (20 b-c) show stem-conditioned (irregular) plural allomorphs, one of which is a zero morpheme, and one of which is an overt suffix, whose distribution is highly restricted. In DM, these irregular plural items are specified for insertion only in the context of a listed set of lexical roots. They are therefore more highly specified than the default item, and thus win the competition for insertion into the syntactic node bearing the plural feature in the context of these roots, ruling out forms, such as *oxes. Inserting one item blocks the insertion of additional items, so forms, such as *oxens are also correctly ruled out.

(20) a. cat ∼cat-s
b. ox ∼ox-en (cf. box ∼box-es)
c. sheep ∼sheep-Ø (cf. beep ∼beep-s)

As pointed out by Grimstad et al. (2014), this Subset Principle plays a vital role in constraining intraword codeswitching. Crucially, during the production of a codeswitch, this allows for phonological exponents from any language to be inserted into the syntactic tree, regardless of the language identity of either the syntactic or phonological elements—assuming the vocabulary item which is inserted meets the demands of the active features in the syntax. In a codeswitching context (i.e., bilingual communicative mode), this allows functional vocabulary items from languages A and B to compete with each other. If the syntactic frame contains features [+X, +Y, +Z], any Vocabulary item matching these features or any subset of these features (i.e., [+X, +Y], [+Y, +Z] or [+X, +Z]) may be inserted into the derivation. However, a Vocabulary Item may not be inserted if it contains additional features not present in the frame (e.g., [+X, +Y, +Z, +A] or [+X, +Y, +B]). This principle will prove crucial to accounting for when intra-language codeswitching may occur.

By separating the syntactic, phonological and semantic components of a word, DM allows for a molecular view of a word which, as we will see in section Producing ICS, is compatible with the MOGUL account[14].

Distributed Morphology and ICS

In a DM account of language mixing, the notion of lexical decomposition also plays a central role in allowing ICS to occur. Lexical decomposition is the notion that category-neutral lexical roots (e.g., *cat, man*, etc.) combine with one or more functional heads in the syntax (e.g., *Gen, Def, Num, etc.*); proponents of lexical decomposition argue that this accounts for complex syntactic meanings (Halle and Marantz, 1994). Notably, in terms of language mixing, it is the construction of syntactically complex words, where a root is from language X and the syntactic features are from language Y, which ultimately allows ICS to occur in DM. This is in contrast to MacSwan's Lexicalist approach, in which morphologically complex words are viewed as syntactic atoms which cannot be syntactically decomposed, thus preventing ICS. As pointed out by Grimstad et al. (2014), when a model of Distributed Morphology is applied to a bilingual/multilingual's lexicon intraword language mixing appears to be part of the natural fallout of language use; this is a sharp contrast to the lexicalist model proposed by MacSwan. In DM, while syntactic trees can only contain syntactic feature bundles from a single language, these frames are blind to the language identity of Vocabulary items. Phonological exponents from either and/or both languages may be inserted into the tree assuming they match the syntactic features of the terminal nodes. That being said, we will still need to explore the issue of whether the vocabulary items from two languages are implemented with two phonologies. In the next section we discuss how to accomplish this objective via experimental methodologies.

[14]This compatibility of DM to both competence and performance accounts of ICS is, for us, a virtue; we are not arguing that it is the only possible theory of morphology which could be implemented in MOGUL.

ICS AT THE REPRESENTATIONAL LEVEL

As made clear in section *Discussion: Morphological Restrictions vs. Phonological Restrictions*, we require experimental data from methodologies specifically designed for the questions at hand in order to provide an accurate account of intraword phonological CS[15]. Our evaluation of the examples in **Table 1** shows that existing research on the phonology of morphological switches largely lacks (a) acoustic information, (b) data from multiple speakers, (c) data from methodologies thought to better tap representation (i.e., phonological perception and acceptability judgment tasks), and (d) comparison of the codeswitching data with data provided in monolingual mode in each of the bilingual's languages (i.e., treating each bilingual as her own control). These comparison data are especially important in light of the individual variation found among bilinguals, particularly in contexts in which one of a bilingual's languages is not the dominant community language. With these four criteria in mind, the data points in **Table 1** are simply not sufficient to support or refute the claim that intraword phonological switches are not possible.

To address the need for experimental evidence, Stefanich and Cabrelli Amaro (2018a,b), Stefanich (2019) have designed multiple empirical measures to tap phonological representation in cases of intraword morphological switches, taking into account the need for both acoustic and judgment data from large sample sizes collected in bilingual as well as monolingual modes. The first phase of the project tested which phonological system(s) a Spanish/English bilingual utilizes in the production of verbs with English roots and Spanish affixes, with the hypothesis that Spanish phonology would be exclusively applied since Spanish is the language of the affixes (see section *Review of*

[15]The only experimental evaluation of potential phonological switches to our knowledge comes from MacSwan and Colina (2014), who examined interword production within English-Spanish mixed DPs, rather than intraword production. They found that phonological processes can apply across word boundaries only if the process is operational in the language of the word that would be affected, an outcome which aligns with the hypothesis that two phonologies do not apply within a word. Akinremi (2017) and Hlavač (1999) explicitly discuss the phonology of mixed words but they do not provide experimental or acoustic data.

Foundational Research on Intraword Codeswitching). Elicited production tasks were completed by 19 English-dominant American English/Mexican Spanish bilinguals who identified as naturalistic code-switchers and had positive attitudes toward CS (see Badiola et al., 2017, for a discussion of effects of attitude on CS behavior). Several design choices were made with the challenges of experimental CS research in mind. Crucially, participants completed three versions of the task on separate days, with the first day always administered in bilingual mode. The bilingual mode session was administered by a member of the participants' bilingual community who was similar in age and was a naturalistic code-switcher. The interlocutor's profile was an important methodological choice, as her presence contributed to a socially motivated codeswitching context (see section MOGUL and Control for discussion of how this type of interaction licenses a bilingual mode). The remaining 2 days of testing were administered in English and Spanish (order was counterbalanced across participants). Testing in these three modes allowed the authors to determine the source of phonology in ICS productions by using the Spanish and English data as a baseline, rather than comparing ICS productions with a monolingual norm (see e.g., Ebert and Koronkiewicz, 2018). Another important consideration was the stimuli design; nonce verbs were used to address the challenges of teasing apart code-switches from borrowings discussed in sections *Review of Foundational Research on Intraword Codeswitching* and *Discussion: Morphological Restrictions vs. Phonological Restrictions* and to control for potential frequency effects. To identify any instances of phonological switching, each English verb in the CS task contained one of three phonemes that are not part of Mexican Spanish (/z/, /θ/, /ɪ/). To provide context during the task, participants were presented auditorily with each nonce verb and a definition and example given in CS. They were instructed to teach the experimenter the new words in "Spanglish" (CS), and prompted in Spanish to produce the verb forms with progressive morphology, the Spanish prompt served to prime participants for a switch into Spanish (21).

(21) Slide 1: *Repite por favor* [please repeat]. To mip.
Slide 2: To mip *es cuando bailas* [is when you dance] to your favorite song in an empty room. *Angela* lives in a studio apartment and she mips every night. ¿*Qué está haciendo en la foto?* [What is she doing in the picture]
Slide 3: Está __________. Expected answer: Está **mip**eando.

The monolingual sessions followed a parallel procedure and served to establish a baseline point of comparison for production of /z/, /θ/, and /ɪ/ in English and predicted Spanish-like substitutions /s/, /t/, and /i/. The authors predicted that, if an English verbal root with Spanish progressive morphology were produced using Spanish phonology (the language of the affixes), bilinguals would not produce English segments in the root. Instead, /z/ was expected to surface as [s], /θ/ as [t] or [s], and /ɪ/ as [i] (e.g., Morrison, 2008; Costa, 2009).

The authors found evidence of application of Spanish phonology across the three phonemes. English /z/ was produced as a Spanish-like [s] in the English root of the mixed word by 50% of the bilinguals. Remaining participants produced [z], which could potentially indicate application of English phonology. However, analysis of these participants' monolingual Spanish production of /s/ in a voicing assimilation context revealed production of [z] in this context. The authors posited that, for these participants, [z] is a part of their Spanish phonetic inventory and thus these data could not serve as conclusive evidence of intraword phonological codeswitching. Data from the /ɪ/ and /θ/ phonemes, which do not have a corresponding allophone in Spanish, back the hypothesis that [z] productions in the English root reflect the Spanish phonetic inventory. Specifically, the bilinguals produced /ɪ/ as [i] in the codeswitching task, which the authors took as an indication that the participants applied only Spanish, and not English, phonology in morphologically switched words. In the case of mixed words with /θ/ in the English root, Stefanich and Cabrelli Amaro (2018b) found evidence of substitution of /θ/ via a handful of Spanish-like sounds, namely [t̪], [s], [z], [v], and [f]. The production data from this first phase of the project therefore is taken as preliminary support for the posited ban on word-internal phonological switches.

The logical question that followed from these data, and which led the authors to the project's second phase, is first mentioned in section *Discussion: Morphological Restrictions vs. Phonological Restrictions*: Although these bilinguals do not produce morphologically switched words with elements from two phonologies, are such structures illicit in their bilingual grammar? With this question in mind, the authors implemented an aural judgment task as a method to tap the participants' I-language by testing the acceptability of phonologically codeswitched words (see e.g., González-Vilbazo et al., 2013, and Schütze and Sprouse, 2014, for motivation for this type of method). The task consisted of morphologically switched verbs with English roots and Spanish progressive morphology; the stimuli were the same as those used in the production task in the codeswitching session. Each trial belonged to one of three conditions: Items produced with English phonology only, Spanish phonology only, or phonology matching the morphology of the item (i.e., English phonology in the root and Spanish phonology in the affixes). To maximize ecological validity, a member of the bilingual population with phonetic training produced all of the stimuli. The items in the phonological switch condition were constructed via splicing of English roots produced in monolingual English mode and Spanish affixes produced in monolingual Spanish mode; doing so ensured presentation of stimuli which exhibited a true phonological switch without phonetic contact effects[16]. Twenty seven bilinguals with the same profile as the participants in the production task completed the judgment task. As an inclusion criterion, participants had to be able to distinguish perceptually between the three conditions for each phoneme; this resulted in the inclusion of data from 24 participants for /z/ and 17 for /θ/ and /ɪ/. The results confirmed the

[16]To control for any potential effects in judgments due to splicing, the authors also constructed the items in the Spanish and English phonology conditions by splicing the root and affixes.

findings from the production task, such that the bilinguals assigned the highest ratings [using a scale of 1 (completely unacceptable/not a possible answer in Spanglish) to 7 (completely acceptable/a possible answer in Spanglish, z-score transformed to account for individual variation in scale use)] to items produced with Spanish phonology (the language of the affixes), lower ratings to the phonologically switched items, and the lowest ratings to items produced with English phonology, which is in line with the hypothesis that the phonology of a morphologically switched word must be the language of the affixes. Together, Stefanich and Cabrelli Amaro's production and judgment data designed explicitly for ICS research provide the first comprehensive experimental account of ICS. The data, while evidencing substantial variation, point toward a trend in which Spanish phonology is employed, and one possible explanation here is that phonological ICS are indeed illicit in Spanish/English bilingual grammars. Replication of the results from this and other language pairings using these and novel methodologies will determine whether the proposed ban on phonological ICS holds up to crosslinguistic scrutiny and further inform the theoretical notions reviewed in sections Review of Foundational Research on Intraword Codeswitching and Discussion: Morphological Restrictions vs. Phonological Restrictions more generally.

Up to this point, our discussion has been limited primarily to language-internal factors and questions of abstract linguistic representation. However, the *use* of codeswitching is the result of an interaction of language-internal and language-external factors; this interaction is the focus of the following section.

PRODUCING ICS

We situate our account of the real-time production of ICS within MOGUL. MOGUL[17] is a modular perspective on language processing presented by Sharwood Smith and Truscott (henceforth Sharwood Smith and Truscott, 2014), with the goal of explaining how language inhabits the mind in real time. The architecture of MOGUL (which stands for Modular Online Growth and Use of Language) draws heavily on Jackendoff's tripartite model of language (Jackendoff, 1997, 2003) where linguistic faculties (i.e., phonology, syntax) are encapsulated modules—in the sense of Fodor (1983)— which interface with motoric and conceptual systems, as well as with general cognitive networks. This architecture is pictured below in **Figure 2**.

Crucially, each module (where PS, Phonological Structure; SS, Syntactic Structure; CS, Conceptual Structure) contains its own unique set of primitive features which are assembled to form representations triggered by linguistic input. For example, SS&T claim that primitive features in the syntax module's information store may include items like [+noun] and [+tense], while the phonological store might contain distinctive features, such as [+strident], [+continuant], or [+voiced]. Once each module constructs its own representation out of the set of primitives available to it, the representation is then interfaced to neighboring modules. The result of this interfacing is a (PS + SS + CS) representational *chain* which contains all the necessary phonemic, syntactic and conceptual information equated with a word. **Figure 3** is an example of a representational chain which would represent what is commonly thought of as the word *lamp* (from Truscott and Sharwood Smith, 2004).

Here we argue that language mixing is the result of constructing representational chains using features from Language-X (Lx) and Language-Y (Ly).

MOGUL, however, is more than just a theory of language or language development; MOGUL is an account of the multilingual mind in which language is plugged into a larger cognitive architecture. The model emphasizes the role of extra-linguistic factors in language production, comprehension and development. The interaction between linguistic and extra-linguistic cognitive systems is particularly relevant to the study of lexical selection (and, more important for our purposes, codeswitching) which may be motivated by any number of non-linguistic factors, such as social circumstances or personal goals. For example, when a speaker is choosing a label to describe members of a militant organization, they will have a number of well-formed options including terms like *freedom fighters*, *rebels,* and *terrorists* (Van Dijk, 1997). In this case, the item that a speaker chooses may be heavily influenced by extra-linguistic such things as their personal experience, emotional reactions, and global politics. Ultimately, the MOGUL platform allows for extra-linguistic aspects of codeswitching (e.g., personal experience) to be accounted for via extra-linguistic processes (e.g., cognitive control).

Basic Architecture

There are two types of modules: linguistic modules and extra-linguistic modules. Linguistic modules include the phonology module (PS) and the syntax module (SS). We assume that PS encodes a phonemic contrastive level of representation (e.g., /kæt/) but the phonetic details of the spell-out would be handled by the production system. Crucially, these two modules are specific to language processing and constitute what we will call the *language core* (see **Figure 2** above). Extra-linguistic modules include the perceptual modules, the motor-control module and the conceptual module. While these modules are involved in language processing (i.e., semantics, speech perception, and production) they are *extra-linguistic* in the sense that they are part of a general cognitive apparatus that governs action and knowledge beyond language. In addition, the CS contains a "general language representation" (GLR); this GLR (e.g., English, Swedish, Yoruba) is triggered by context and co-indexed with all representations associated with the language it represents. This point will prove crucial when accounting for language selection in the MOGUL framework and will be expanded upon in section Cognitive Context, Conceptual Triggering, and Language Selection.

Information stored in each module includes what SS&T call an *activation level*. When static, content in the information store will sit at a resting level of activation. All elements will have a resting level of activation based on previous usage as well as the strength of associations to other units. The activation level

[17] MOGUL has become part of a larger research program known as the Modular Cognition Framework (MCF). In this paper, we will still use the term MOGUL. For more information, see the website: https://www.cognitionframework.com.

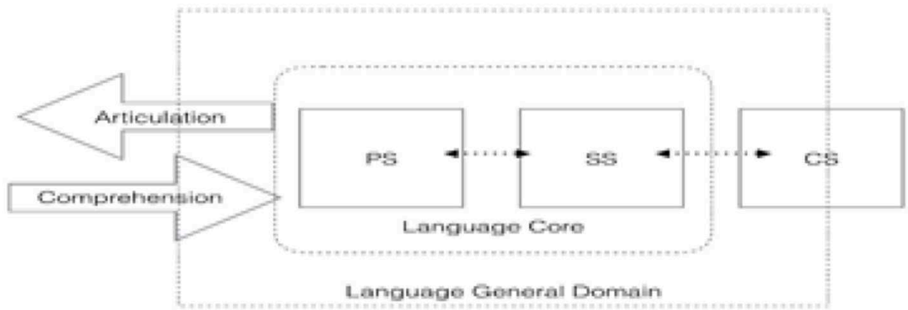


FIGURE 2 | Tripartite model (adapted from Sharwood Smith and Truscott, 2014).

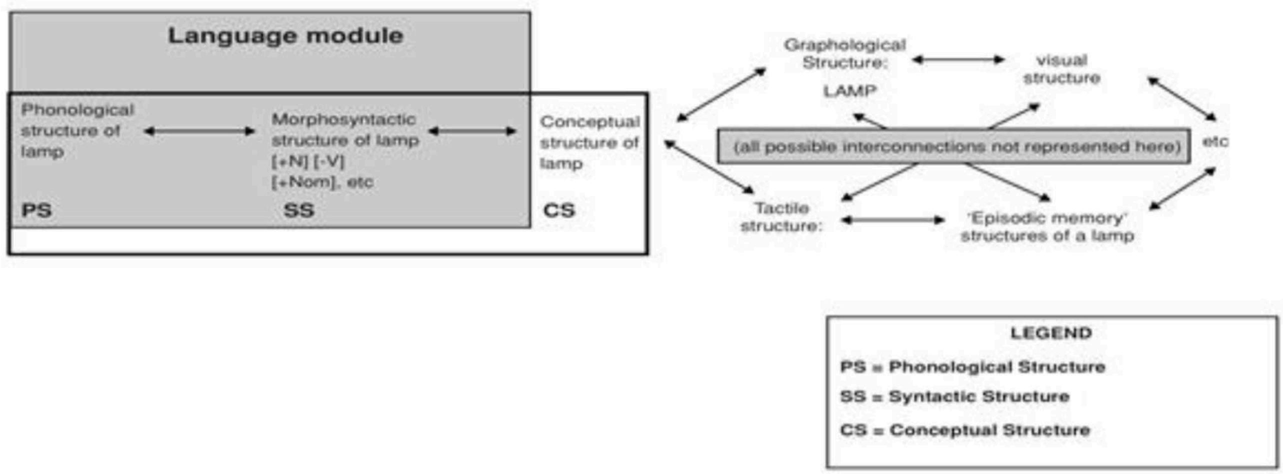


FIGURE 3 | A co-indexed representational chain.

of an element in the information store is affected by *spreading activation* between associated elements (Rumelhart et al., 1987). For example, Sharwood Smith and Truscott (2014) note that a listener's interpretation of the term *bank* will be affected by whether or not they just heard the term *river* or *money*. This occurs as the terms *river* and *bank* each prime a specific meaning of the term *bank*—that is, either, the side of a river, or a secure location to deposit one's money. However, as the two senses of the term *bank* are homographs and homophones, they also prime each other (Sharwood Smith and Truscott, 2014). So, the term *bank* will in isolation cause the processor to activate both senses of the word, but collocation with another term, such as *money*, will increase the activation level of a particular sense of *bank*—in this case, let us say a financial institution—which will cause it to win the competition.

But, as Green (2018) demonstrates, we need more than activation to account for the production of a mixed utterance; we need to include a theory of selection in the production planning process.

Of course, language is more than a grammar. It is a vehicle which allows us to reveal our inner thoughts and feelings to others. It is a tool which allows us to elucidate the pain we are experiencing or the lovely mountains we are seeing while talking on the phone to someone far-removed from the visual stimulus. Let us now expand the model slightly to introduce the sensory and emotional interfaces.

The core language modules interface with the sensory domains of cognition known as Perceptual Output Structures (POpS). Two of these modules (audio and visual structures) are most relevant to language processing (as shown by the connecting lines in **Figure 4**), but all five sensory-perception modules (including gustatory, olfactory and tactile) can interface with the language core.

MOGUL also recognizes the emotional aspects of language use via an affective module—labeled AffS. Similar to POpS, the AffS is a part of general cognitive processing and not specific to language. These AffS structures assign a value feature (e.g., +1 or −1) to interfaced (i.e., co-indexed) representations; higher values in the representations correspond to elevated levels of activation for representations that are part of the same chain. As SS&T explain, "These AffS structures have the effect of assigning either a positive value or a negative value to the

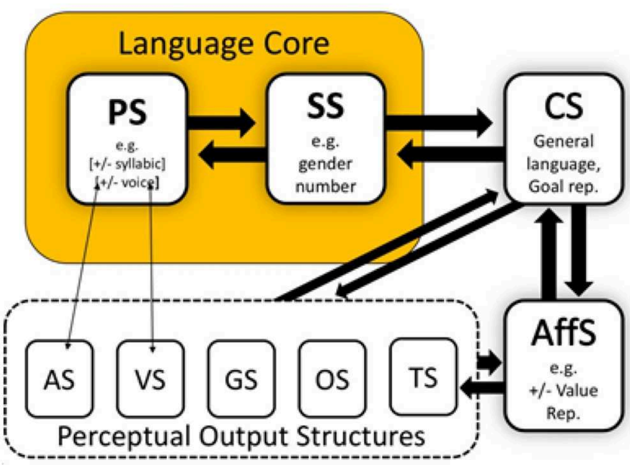


FIGURE 4 | An elaborated model.

representations they are co-indexed with" (2016: p. 3). A lexical example should clarify. Consider the words *cheap* and *thrifty* in the context of "Gareth is ____." Conceptually, these terms are viewed as nearly synonymous, differing only in that *thrifty* has a positive value representation in the AffS, while *cheap* has a negative one. Thus, the semantic chain of representations (i.e., CS + POpS + AffS) for these two lexical entries may be nearly identical, differing in that *thrifty* would be represented as [positive] in the AffS while *cheap* would have the [negative] AffS representation. As such, the selection of one term over the other may be emotionally motivated; the choice of whether Gareth is *cheap* or *thrifty* may depend on whether or not the speaker likes Gareth.

Together, the POpS (perceptual) and AffS (affective) systems play a major role in establishing lexical knowledge. Our conscious understanding of a word's meaning is the synthesis of our perceptual understanding, affective reaction, and CS. More technically, MOGUL accounts for lexical meaning via a (CS + POpS + AffS) chain of representations. A lexical entry is not a single representation but rather an amalgamation of module specific representations that are both external (CS + POpS + AffS) and internal (PS + SS) to the language module. Specifically, a lexical entry is a chain of co-indexed representations [PS + SS + CS (POpS + AffS)] (see **Figure 3**). The lexicon in MOGUL is merely a subset of highly structured long-term memories which contain patterns of activation (i.e., co-indexations) for feature bundles in multiple modules. In other words, the lexicon consists of chains of representations (i.e., lexical entries) stored in long-term memory. This is one of the reasons why MOGUL is able to implement a Distributed Morphology model in which morphological knowledge is distributed across various representational modules. In sum, conceptual representations and extra-linguistic knowledge are central elements in a MOGUL account of codeswitching; these elements work together to form lexical meanings in the form of complex CS representations.

MOGUL and Cognitive Context

Clearly, in production, certainly at the word and sentence level, bilinguals have conscious control over which language they choose to speak. A central element in accounting for language selection in MOGUL is *cognitive context*. The notion of cognitive context stems from work on mental models in the field of cognitive science (Johnson-Laird, 1980; Van Dijk, 1997) and may generally be characterized as a mental model which an individual creates to reflect their environment. Such a model is heavily influenced by personal experience (e.g., personal perceptions, pre-conceived opinions, etc.). Factors which may influence cognitive context include (Van Dijk, 1997: p. 193):

- Setting: location, timing of communicative event;
- Social circumstances: previous acts, social situation;
- Institutional environment;
- Overall goals of the (inter)action;
- Participants and their social and speaking roles;
- Current (situational) relations between participants;
- Global (non-situational) relations between participants;
- Group membership or categories of participants (e.g., gender, age).

All of these factors situate an individual in a particular environment. Taken together, this is cognitive context. When discussing codeswitching, cognitive context supplies the set of mental representations that lead an individual to produce one language instead of another in a given setting. SS&T note that an increased activation of "particular languages [(that is PS + SS)] will… be triggered by given patterns of existing conceptual and affective structure: the initial source may be either internal or located in the observable environment" (Sharwood Smith and Truscott, 2014: p. 198). In other words, language selection in production is cued not just by linguistic input being processed but by the social, political and environmental factors emanating from the real-world situation an individual is experiencing (i.e., external factors, such as street signs) as well as an individual's perceptions of said experience (i.e., internal conditions, perceived language prestige).

Influences on Cognitive Context

Three key factors have been identified in playing a critical role in establishing a context which is conducive to language mixing. These factors are:

1. Linguistic Landscape
2. Identity of Language User/ Self-image
3. Identity of Interlocutor

Let us consider them each in turn. An individual's linguistic landscape is a central external factor in establishing cognitive context (Sharwood Smith and Truscott, 2014). A linguistic landscape refers to the real-world linguistic environment individuals find themselves in, taking into account visible aspects of language in a given environment, everything from road signs to gestures. This linguistic landscape is then mentally internalized and becomes part of cognitive context. For example, an individual witnessing a number of English street signs, as

FIGURE 5 | Street signs contribute to a language user's linguistic landscape.

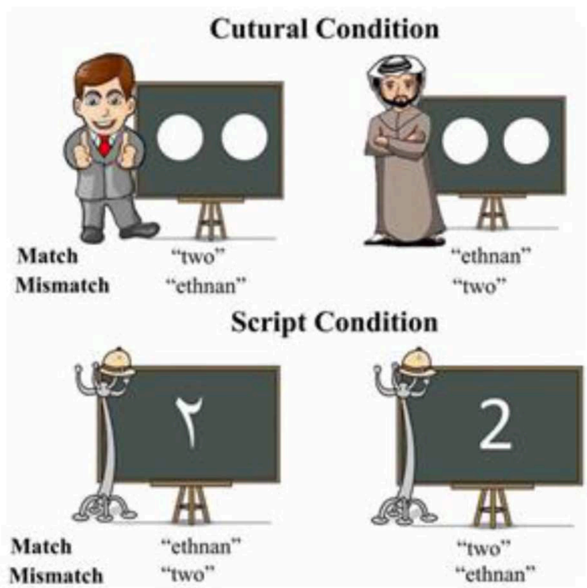


FIGURE 6 | Stimuli from Blanco-Elorrieta and Pylkkanen (2015: p. 2).

pictured in **Figure 5**, would form an English linguistic landscape which would promote English productions.

The second factor is the language user's self-representation (i.e., a language user's view of *self* in their cognitive context). While, crucially, the language user must self-identify as a bilingual (or multilingual), self-representation also refers to a language user belonging to or identifying with various social/cultural groups (e.g., university students, musicians, cognitive scientists, etc.; Van Dijk, 1997). The identity (or perceived identity) of the interlocutor is the third factor. The language user must believe that the interlocutor is a bilingual in order to license language mixing. Additionally, if the language user and the interlocutor identify with similar social or cultural groups, the language user may perceive an affinity with the interlocutor which sets the stage for socially motivated codeswitching, spurred by notions of solidarity and group belonging (Poplack, 1980).

Cognitive Context, Conceptual Triggering, and Language Selection

A fundamental property of bilingual performance is a speaker's ability to selectively produce one language or another without mixing elements cross-linguistically. Before we can understand how two linguistic systems can be brought together, we must first examine how they are separated. Thus, any account of language mixing necessitates an account of language *selection*. This account of language selection is motivated by the observation that bilinguals tend to use specific languages in specific contexts (e.g., language X is used at work but language Y is used at home) (Truscott and Sharwood Smith, 2016). Within the MOGUL framework, a rich cognitive context functions to activate elements in each module's information store that are associated with a specific situational context. This process is known as *conceptual triggering* (Sharwood Smith and Truscott, 2014). Conceptual triggering relies on a specific CS feature called a *general language representation* (GLR) which increases the activation level of co-indexed representations associated with a specific language variety. SS&T state, "the contexts that have a triggering effect are particular perceptual and conceptual structures that are associated with elements of one of the person's languages, including general language concepts like FRENCH or YORUBA (2014, p. 199). However, it is also possible that GLRs could activate particular varieties or registers of a language, such as *formal* French or *informal* Yoruba. These general language representations are CS representations co-indexed to feature bundles associated with the relevant language variety in the PS and SS; when a GLR from language X becomes active in the CS, all representations which are co-indexed to X-GLR receive an activation boost to their current level of activation. It is worth noting that linguistic representations may be co-indexed to multiple GLRs, as in the case of cognates and interlingual homophones.

This machinery can clearly be employed to account for intentional language mixing at the lexical, phrasal, or sentential level, but we will return later to the question of whether intraword codeswitching is intentional.

MEG studies by Blanco-Elorrieta and Pylkkanen (2015, 2016) have investigated the role of natural occurring language cues (i.e., ethnicity of the interlocutor and orthography) in triggering language selection. These language cues are factors involved in establishing cognitive context. In their study, Arabic-English speakers perform two number-naming tasks: a match task and a mismatch task. For each task there were two conditions, a script condition (i.e., orthographic) and a cultural condition (i.e., a culturally iconic picture). These conditions are displayed in **Figure 6**.

During the match task, participants were shown an image and asked to name the number indicated on the blackboard in the language indicated by either the script or the clothing. The mismatch task required the opposite: participants were to name the number in the language which was not cued by the script or clothing. The results of this study revealed that script—which is part of a linguistic landscape—was a much more effective cue than the cultural condition in terms of triggering language selection.

Neurologically, this distinction between the cultural and orthographic condition is demonstrated by a greater degree of activation in the Anterior Cingulate Cortex (ACC) during the cultural condition (compared to the script condition) which they attribute to greater processing difficulty. Prior research has implicated the ACC as having a major role in establishing attention (Abutalebi et al., 2008, 2013; Costa and Sebastián-Gallés, 2014) which, in turn, is related to executive control (Blanco-Elorrieta and Pylkkanen, 2015: p. 14).

According to Blanco-Elorrieta and Pylkkanen (2015, p. 14), executive control is "the effort to retrieve…a word amongst competing responses" during language production. As several studies have reported an increase in ACC activity for greater executive control demands (Abutalebi et al., 2008; Garbin et al., 2011; Costa and Sebastián-Gallés, 2014), they suggest that the increase in ACC activity during the cultural condition reflects a greater effort to retrieve the target element; the cultural condition is a weak language cue, and requires additional cognitive effort to satisfy production goals. Alternatively, the lesser demand on the ACC during the script condition seems to indicate that script is a strong cue which appears to dominate cognitive context when selecting a language.

MOGUL and Control

Researchers have often turned to the notion of executive control to explain language control (Green, 1998; Abutalebi et al., 2015). Among its many functions, an executive control mechanism allows language users to suppress representations from language A while permitting representations from language B to surface. Early models of executive control proposed single mechanisms which were capable of performing a number of quasi-related functions which were otherwise unaccounted for. For example, Green (1998) discusses a control mechanism called the Supervisory Attentional System which, "… must command a variety of processes, including the construction or modification of existing schemas and the monitoring of their performance with respect to task goals" (Green, 1998: p. 69). However, more recent work (Abutalebi et al., 2015; Green and Kroll, 2019) recognize the complexity of these extra-linguistic processors. While space does not permit us to explore details of these domain-general control mechanisms, it is uncontroversial to note that there is strong neurolinguistic evidence for the distinction between what Green and Kroll (2019) call the language network and its control.

This is reflected in the work of Green and Wei (2016) on codeswitching and language control. They argue that codeswitching is not the product of activation within the language system, but rather the result of selection in the planning process. When one language is suppressed entirely in production, this is the result of competitive control. However, when forms from more than one language are selected, this is the result of cooperative control. There are two ways in which this can happen. First is what they call *coupled* control. Under coupled control there is a dominant, target language of the utterance even though items from another language may be inserted. Second is what they call *open* control. Open control licenses dense codeswitching. Under this scenario, there is no dominant, target language but rather both languages are active and the planning process can select items from either. In this framework, the different types of output are associated with different attentional states. Dense codeswitching would be associated with a broad attentional state[18].

From a MOGUL perspective, where modules are believed to be "expert systems" which perform specific tasks, there is no reason to believe that a single mechanism is responsible for performing such seemingly unrelated functions as goal maintenance, schema construction, or conflict monitoring. MOGUL embraces a pluralistic construct of executive control which is consistent with that of Green and Wei (2016). "[There is] no single fixed executive control but [instead] different mental subsystems operate in a way that may be highly constrained" (Sharwood Smith and Truscott, 2014: p. 21). As such, the functions of executive control are broken down and attributed to specific modules. Of the many functions attributed to executive control (e.g., goal maintenance; conflict monitoring; interference suppression; salient cue detection; selective response inhibition; task disengagement; task engagement; opportunistic planning), the notion of *goal* formation and maintenance is central to our account of ICS in MOGUL.

Goals are realized via goal representations which are constructed in the CS and interfaced to neighboring modules. To illustrate this, let's further explore the notion of goal. Truscott and Sharwood Smith (2016) claim that goal representations are a type of CS structure which help guide thought and action. Goal representations "serve the function of encouraging the satisfaction of basic needs" (Truscott and Sharwood Smith, 2016: p. 5); these basic needs may be non-linguistic and could include the desire for food, water, bathroom, etc. These goals motivate action (Damasio, 2018) such that goals with higher activation levels are prioritized and drive an individual to perform tasks which satisfy the goal. For example, a basic goal like "satisfy hunger" can be satisfied by eating.

In the realm of language production, goal representations (which are most likely below consciousness) related to communicative and social functions are particularly pertinent. SS&T argue that such social goals are formed from a set of primitive CS features representing social motivators which could include *affiliation*, *power* or *face*. Language-oriented goal representations serve to motivate language use. Goal representations play a significant goal in determining the shape and style of language production by increasing the activation level of linguistic representations which can satisfy the goal. These goal representations are co-indexed to value representations in the AffS; the higher the value in the representation, the higher priority the goal, and hence the greater the increase in activation levels to linguistic representations that satisfy the goal.

Let us imagine the following scenario to understand how this machinery can work in a real-world situation. For narrative

[18]Green (2018) further explores the constructs of coupled and open control. In MOGUL terms, the multilingual cognitive context (and the likely dense codeswitching input) appears to prime open control. Given the results of Stefanich (2019), it would be worth exploring whether the morphological and phonological switches are governed by coupled or open control. The lack of phonological switching could perhaps be the result of the short timespans involved.

purposes, we draw on the population which is the source of the Alexiadou et al. (2015) Norwegian/English data: a group of people of Norwegian descent who live in the United States. Let us return to our hypothetical group member Gunnar. When Gunnar is interacting with monolingual English speakers in Minnesota, he usually produces English-only utterances, suppressing the production of Norwegian. In situations, though, where the linguistic landscape includes such things as other Norwegian speakers, or Norwegian food products, or literature, his speech may contain language mixing at the sentence-, phrase-, or word-level. These external factors influence the cognitive context which licenses a bilingual communication mode. When in this mode, Gunnar may have a variety of social goals including *let them know I'm Norwegian too*, or *reminisce about the summers at the lake when I was a child* or *imitate the funny way that a certain politician spoke*. Codeswitching, then, might be a vehicle to satisfy any of these social goals. Linguistically, the tools used to achieve these goals could be a codeswitch within an utterance, or a sentence, or a word. Codeswitching is not the goal.

MOGUL and Communicative Modes

When discussing bilingualism, a number of language mixing researchers have suggested that bilinguals are able to exploit different *modes* of communication (Poplack, 1980; Poplack et al., 1988; Myers-Scotton, 1992; MacSwan, 1999; Grosjean, 2001; Sharwood Smith and Truscott, 2014). Communication modes are invoked to account for a bilingual's ability to generally suppress one language in production as appropriate to the social setting. In MOGUL, these communicative modes are cognitive states where contexts and goals align to produce the contextually relevant language. When a speaker engages in a bilingual mode, representations from two languages are equally active in the production process. This would be consistent with Green's notion of open, cooperative control. It is this configuration of contextual representations that licenses the mixing of grammatical systems during codeswitching (including ICS). Both external (e.g., linguistic landscape) and internal factors (e.g., goal representations) contribute to the set of conditions which make up a bilingual context; a context conducive to language mixing. The goal representation is, however, crucial for providing the language user with the *motivation* for engaging in a codeswitch. While these motivations vary—from pride to humor to solidarity—it is goal representations which drive a language user to produce an ICS.

Following Grosjean (2001) we will adopt the term *bilingual mode* (as opposed to mixed speech mode or codeswitching mode). In such circumstances, it is the *act* of switching, not the location of switch, that may carry meaning— if the goal of the utterance (or the act of codeswitching) is to promote solidarity, the type of elements switched may not be relevant. A bilingual may switch a consonant (saying *Bach* as [bax] or [bak]), a word (*careful!* or *¡cuidado!*), or a morpheme (*den field-a* "that field") and it signals to the interlocutors that they are members of the same social group.

So, while the production by the speaker may not be a conscious decision to switch at a particular point, the bilingual mode (or open control) licenses the switch. The listeners then recognize the intended goal (e.g., group solidarity).

We saw earlier that representational chains are constructed from the co-indexation of module-specific feature bundles (e.g., PS + SS + CS). What we introduce now is the fact that module-specific features bundles associated with different languages can come together, resulting in an ICS.

When the bilingual communicative mode is engaged, *two* languages are conceptually triggered and all representations co-indexed to *either* GLR receive an activation boost. Crucially, goal representations are co-indexed to any representation which helps satisfy that goal which, in turn, increases the activation level of said representation. A representational chain is then formed from the most active content in each module; the resulting chain will contain a SS representation from, say, Lx, a PS representation from Ly and CS representation from both Lx and Ly—this is an intraword codeswitch, as shown in (21).

21) den **field**-a
that field-DEF.F
"that field"

Notably, there are no special processes or mechanisms evoked to account for codeswitching; language mixing is the natural product of standard MOGUL operations and processes. As Truscott and Sharwood Smith (2016: 903) note, somewhat paradoxically, "a theory of codeswitching should, ideally, not be a theory of codeswitching." In this model, codeswitching can result from a particular cognitive context and a bilingual communicative mode. The bilingual mode is a reflection of cognitive context, while codeswitching is the act which *satisfies* a particular communicative goal. When an individual produces an ICS, they do so because bilingual mode is active and codeswitching is the best way to satisfy their goal.

Grammatical Machinery

This account of codeswitching production begins with cognitive context. An internalized mental model is constructed by the speaker to reflect the external environment they are experiencing. We will refer to the language which is conceptually triggered as the *prominent* language. We do so to avoid confusion with the term *dominant* which is used in the Matrix Language Frame Model. All representations which are associated (i.e., co-indexed) with the triggered language will receive an additional activation boost. This activation boost means representations co-indexed to the conceptually triggered language will usually come out on top in any competition.

However, in language-mixing situations, cognitive context is oscillating between two prominent language contexts. This notion of oscillation is a crucial one; only one language can be prominent (i.e., conceptually triggered) at a given time; however, context changes in real time. These changes may be external (i.e., changes to the physical environment) or internal (i.e., re-evaluating goals). The metaphor of contextual oscillation represents a rapid alternation of prominent linguistic contexts. This fluctuation between contexts can, in principle, happen a number of times over the course of constructing a single word, or phrase, or sentence. The word level is not a special barrier to language mixing. Such a view is consistent with a dense codeswitching environment which can trigger open cognitive control.

When a speaker experiences a cognitive context which conceptually triggers the prominent language, the result is the construction of a syntactic structure as the speaker plans an utterance. It is the nature of this tree, and the single engine which drives both the morphology and syntax which we focus on here.

In DM, once the syntactic tree has been constructed, Vocabulary items compete for insertion. In order for intraword codeswitching to occur, a speaker must engage the bilingual mode of communication. DM appears to be highly compatible with MOGUL in that DM vocabulary items correspond to PS representations and their co-indices to SS representations. Under most circumstances, the PS representation inserted into the syntactic tree is the PS representation which best matches representations in both the SS and the CS, and which has been conceptually triggered. We are not claiming here that MOGUL requires a DM architecture. Sharwood Smith and Truscott (2014) note that MOGUL is a *framework* which can implement many different formal models. The goal of this paper is not to assess the broader falsifiability of MOGUL.

When a speaker is in bilingual mode, representations from both languages are able to compete with each other. A DM framework would allow for Vocabulary items from both languages to compete to be inserted into the syntactic tree. Within the MOGUL framework, the activation level of any PS representation will be determined by the activation levels of all co-indexed representations in other modules. If an SS feature bundle increases in activation, representations with co-indexed features will also increase in activation; the PS representation with the most features co-indexed to the active SS representation will be the most active in the PS module and will be inserted into the derivation; this is competition in MOGUL.

During the processing of a derivation in MOGUL, lexical items are formed from the most active co-indexed representations. When considering an instance of codeswitching, each feature bundle in the SS is potentially co-indexed to a pair of PS representations; one from the language that serves as the host in that moment and one from the donor language. In a language mixing situation, a speaker has two general language representations active in their CS thus conceptually triggering two languages. Representations co-indexed to either GLR will receive an additional activation boost. Effectively, this means that PS representations associated with two different languages are competing on an equal playing field; from here competition may proceed in the standard MOGUL fashion.

However, the question remains: when elements from the prominent and attenuated language are in competition, what allows the attenuated language to ever win? In standard bilingual environmental circumstances (i.e., oscillating contexts), when PS representations from both languages are in competition, one would predict that the prominent representation should always win the competition as frequent, salient representational chains will have higher resting levels of activation than newly formed chains; even on a level playing field the prominent language should still win. What are the circumstances under which the attenuated language could ever win? We argue that this phenomenon can be accounted for via the construct of goal representations. Communicative goals contribute to cognitive context but also increase the activation level of representations which satisfy their demands. When a speaker engages a bilingual mode, communicative goals conducive to codeswitching will be active in the CS. In a language mixing situation, this translates to an increase in the activation level of the attenuated language representations in order to satisfy the goal. As such, goal representations play a central role in permitting a speaker to mix languages.

Let us return to the example drawn from a speech corpus of heritage Norwegian speakers in the USA, repeated in (22):

22) den **field**-a
that field-DEF.F
"that field"

Example (22) is a phrasal constituent which is presumably part of a larger sentence or conversation. The prominent language—conceptually triggered via a specific context—is assumed to be Norwegian, as evidenced by the fact that the syntactic structure was built from Norwegian features. This can be seen by the fact that such mixed sentences show Norwegian V2 syntax. The sentence in (23) is reported in Alexiadou et al. (2015).

(23)	Så	**play**-de	dom **game**-r
	Then	play-PAST	they game-INDEF.PL

"Then, they played games."

This sentence with clear English lexical items is inserted into a syntactic structure which has the verb in second position (V2) as is standard in Norwegian. This cognitive context will reflect a number of internal (e.g., self-image, inter-personal relationships, etc.) and external influences (e.g., location, identity of interlocutor, etc.) which the speaker has co-indexed with the Norwegian GLR in the CS. This means that all modular representations which are interfaced to the Norwegian GLR will receive an activation boost that subsequently causes these representations to dominate any competition. As such, when a speaker chooses to speak, a syntactic tree will be constructed from feature bundles associated with Norwegian.

Let us note here, that while MOGUL includes representational components, it is primarily a model of performance. As a production model, it cannot *predict* which switches will occur in a given utterance *a priori*. It can only work after the fact to attempt to account for why the switches occurred the way they did. In this light, we do not feel that our account of performance falls into a trap of circularity of the following sort:

Q: Why did they codeswitch?
A: Because they had a goal of codeswitching.
Q: How do you know they had a goal of codeswitching?
A: Because they codeswitched.

The goal of a production model is not to predict which utterance will occur. In our view, this does not diminish the contribution of the model in any way. MOGUL, although it does not allow us to distinguish, say, well-formed ICS strings from ill-formed ICS strings, does add clarity to how real speakers use real languages in real situations.

To illustrate how this model of DM accommodates language mixing let us return to the Norwegian-English example introduced in earlier. We propose that in order for ICS to occur, the speaker has to engage a bilingual communication mode and, as a result, codeswitching may serve the communicative goals. The analysis presented in this paper will closely follow work by Grimstad et al. (2014) and Alexiadou et al. (2015). These data were collected from an American heritage community of Norwegian-English speakers and was drawn from the CANS (Corpus of American Norwegian Speech), which is a spoken corpus. Remember that goals are part of a complex CS representation; any representation which is co-indexed to a goal receives an additional activation boost. High-priority goals are co-indexed to high-value representations in the Affective Module (AffS); the higher the value stored in the representation in the AffS the greater the activation boost spreading through the system. The high-priority goal will result in an activation-level boost to co-indexed vocabulary items. We propose that high-priority goals are further *enhanced*. This enhancement provides a greater increase to activations levels than regular co-indexed representations; in **Figure 7** enhanced representations are marked with (+) to indicate an additional boost (i.e., 4+ > 4). As the English PS representation is co-indexed to a goal it receives an activation boost of (1+) which raises its activation level to (4+) and makes it the most active PS feature bundle representing the root FIELD. The most active features in each module are then chained together and result in the English-Norwegian codeswitch, *den field-a* as seen below in **Figure 7**.

To illustrate, we model this process in **Figure 8**. Starting on the right-most side of **Figure 8**, contextual factors are associated with language; some factors like the identity of the interlocutor or a self-representation may be co-indexed to multiple languages—this is represented by multi-colored contexts.

This mixed language context causes the conceptual triggering of two languages in a near simultaneous (or oscillating) fashion which in turn allows representations from English and Norwegian to compete against each other. In **Figure 8**, the bolded circles represent feature bundles which are the most active in their respective modules; the most active feature bundle in each module is interfaced to form a (PS + SS + CS) representational chain. The result is ICS; the formation of representational chains which contain feature bundles associated to two different languages.

The picture of the Lexicon being sketched out in MOGUL looks quite different from traditional lexicalist approaches in generative linguistics (e.g., MacSwan, 2005). Recall that in MOGUL, a lexical entry is not a single representation but rather an amalgamation of module- specific representations that are both external (CS + POpS + AffS) and internal (PS + SS) to the language module. Specifically, a lexical entry is a chain of co-indexed representations [PS + SS + CS (POpS + AffS)]. Thus, the MOGUL architecture is consistent with the performance of codeswitching (including ICS) via vocabulary insertion into a syntactic structure to satisfy the goals of the speaker situated in a particular linguistic and cognitive context.

CONCLUSION AND FUTURE DIRECTIONS

In this paper we hope that we have documented the widespread nature of the phenomenon of ICS. Switching languages within a word is a property of bilingual speech invoking many different languages. And yet, in order to understand this real-world phenomenon, to understand how people like Gunnar and Fulana communicate effortlessly and successfully in a multilingual environment, we need to have sophisticated models of both linguistic knowledge and linguistic performance. In our view, a model of Distributed Morphology couched within the tenets of Minimalist Syntax provides the descriptive and explanatory power to account for the observed facts of ICS. Bilinguals know that ICS is possible but governed by the same type of grammatical machinery that we see in monolinguals. This grammatical machinery is, of course, not open to conscious introspection but nonetheless our analysis demonstrates the principled nature of intraword codeswitches.

Our review of 57 different codeswitched words reveals that the morphological and phonological restrictions on combining elements from two languages in a single word differ from one another. Bilinguals can easily combine roots and affixes from two different languages to form a codeswitched word, and they do so in a systematic manner. While roots can come from L_A and affixes from L_B, it is generally thought that all affixes in a mixed word must come from a single language. However, our crosslinguistic review of ICS suggests that targeted, experimental research is needed in order to confirm that this is the case. Further, for any given language pair, there seems to be a directional asymmetry with respect to ICS such that the root may come from L_A and the affixes from L_B, but not the reverse (i.e., root from L_B and affixes from L_A).

Contrary to mixing roots and affixes from two languages in a single word, it does not seem to be the case that bilinguals can easily mix elements from two phonological systems within a single word. That is, even though a root may come from L_A and affixes from L_B, the sounds that a bilingual uses to utter this word come from a single language. Our review of ICS indicates a prevailing pattern-the phonology of a codeswitched word matches that of the language of the affixes. We reviewed two experiments (Stefanich, 2019) designed to explicitly test this observation. The results indicate that, while there was considerable variation, in general, Spanish/English bilinguals produced words with English roots and Spanish affixes with Spanish phonology. Future work should examine the phonology of ICS via different language pairs, different types of words (e.g., nouns vs. verbs), and different acoustic cues. Additionally, future research should also examine ICS in its larger context, i.e., beyond the word, to see if any additional patterns can be found with respect to how factors, such as linguistic background of the interlocutors, sociolinguistic context of the discourse and syntactic context of the mixed word inform our understanding of the morphological restrictions and phonological restrictions on ICS.

Bilinguals also have not just knowledge, but ability. Linguistic competence can be *used* to signal solidarity, channel identity, and achieve any one of the myriad goals that any speaker can seek to satisfy. Linguistic performance integrates seamlessly with other cognitive faculties as we move through our everyday lives, talking, listening, planning, assessing, looking, touching,

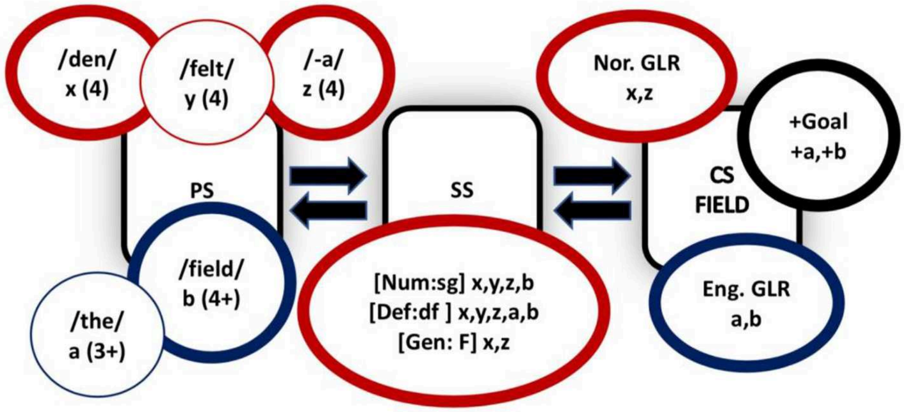


FIGURE 7 | Feature co-indexation with oscillating context and a goal. Bold, most active representation; red, Norwegian; blue, English; round square, module; circle, feature bundle.

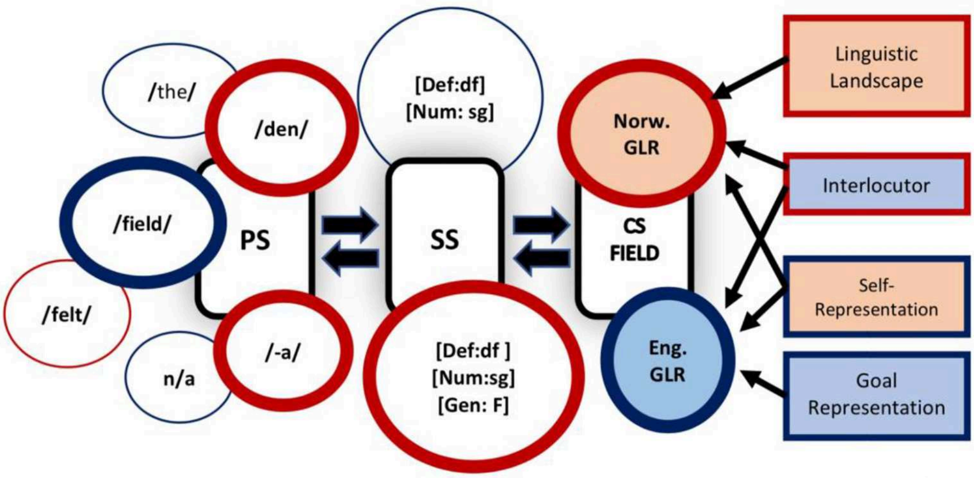


FIGURE 8 | Derivation for *den fielda*. Bold, most active; red, Host/Norwegian; blue, Donor/English; round square, module; circle, feature bundle; rectangles, contexts.

feeling, thinking. In our view, the MOGUL framework provides the necessary cognitive machinery to account for the real-time instantiation of ICS.

Competence and performance models seek to answer different questions and the questions surrounding ICS have just started to be asked. Future methodologies to further probe the character of ICS should draw on a combination of corpora and experimental research. Corpora can be a good starting point, as they give examples in real world contexts. However, we may be unable to answer certain technical questions from corpora alone (such as whether morphological switch boundaries coincide with phonological switches). We must remember though that just because a particular phenomenon is not found in a corpus, it does not mean the phenomenon is ill-formed in competence – it may have been problematic given some performance domain. Thus, assessing well-formedness via acceptability tasks can also advance our understanding of the phenomenon.

A performance model will never be able to predict with absolute certainty the form an utterance will take. As the philosopher Alfred North Whitehead once said to B.F. Skinner at a dinner, no behaviorist model of speech is able to predict someone saying *There is no black scorpion falling on the table*. The structural properties of intraword codeswitches are not fully understood. We are not able to predict whether Gunnar will say *the* field or *den fielda*, or whether Fulana will say *hangear* or *hang out*. But now that we are asking questions about *why* the switches occur where they do and why *not* in other places, why certain structures are allowed, and why others are not, we have a program of research that has the possibility of finding answers.

What's in a bilingual word? Only morphology, syntax, concepts, sounds, features, recursion, connotations, and oscillating contexts. A single word can reveal so much.

DATA AVAILABILITY STATEMENT

All datasets generated for this study can be requested directly from the first author: sara.stefanich@northwestern.edu.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of University of Illinois at Chicago Internal Review Board with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the University of Illinois at Chicago Internal Review Board.

AUTHOR CONTRIBUTIONS

SS and JC were the primary contributors of the linguistic competence section. DH and JA were the primary contributors of the linguistic performance section.

FUNDING

The data presented in Section ICS at the Representational Level was collected as part of a larger research project that was funded by the National Science Foundation Doctoral Dissertation Research Improvement Grant, BCS #1823909.

REFERENCES

Abutalebi, J., Annoni, J.-M., Zimine, I., Pegna, A. J., Seghier, M. L., Lee-Jahnke, H., et al. (2008). Language control and lexical competition in bilinguals: an event-related FMRI study. *Cereb. Cortex* 18, 1496–505. doi: 10.1093/cercor/bhm182

Abutalebi, J., Canini, M., Della Rosa, P. A., Green, D. W., and Weekes, B. S. (2015). The neuroprotective effects of bilingualism upon the inferior parietal lobule: a structural neuroimaging study in aging Chinese bilinguals. *J. Neurolinguistics* 33, 3–13. doi: 10.1016/j.jneuroling.2014.09.008

Abutalebi, J., Della Rosa, P. A, Ding, G., Weekes, B., Costa, A., and Green, D. W. (2013). Language proficiency modulates the engagement of cognitive control areas in multilinguals. *Cortex* 49, 905–911. doi: 10.1016/j.cortex.2012.08.018

Acquaviva, P. (2009). "Roots and lexicality in distributed morphology," in *Fifth York-Essex Morphology Meeting (YEMM)* (York: Department of Language and Linguistic Science; University of York).

Akinremi, I. (2017). Phonological adaptation and morphosyntactic integration in Igbo-english insertional codeswitching. *J. Univ. Lang.* 17, 53–79. doi: 10.22425/jul.2016.17.1.53

Alexiadou, A. (2017). Building verbs in language mixing varieties. *Z. Sprachwiss.* 36, 165–192. doi: 10.1515/zfs-2017-0008

Alexiadou, A., and Lohndal, T. (2017). "The structural configurations of root categorization," in *Labels and Roots*, eds L. Bauke and A. Blumel (Germany: de Gruyter Mouton), 203–232.

Alexiadou, A., and Lohndal, T. (2018). Units of language mixing: a cross-linguistic perspective. *Front. Psychol.* 9:1719. doi: 10.3389/fpsyg.2018.01719

Alexiadou, A., Lohndal, T., Åfarli, T. A., and Grimstad, M. B. (2015). "Language mixing: a distributed morphology approach," in *Proceedings of NELS*, *Vol. 45,* eds T. Bui and D. Özyildiz (Cambridge, MA), 25–38.

Amuzu, E. K. (1998). *Aspects of grammatical structure in Ewe-English codeswitching* (Unpublished MPhil thesis), University of Oslo, Oslo, Norway.

Arad, M. (2003). Locality constraints on the interpretation of roots: the case of Hebrew denominal verbs. *Nat. Lang. Linguist. Theory* 21, 737–778. doi: 10.1023/A:1025533719905

Archibald, J. (2016). *Phonology at the Interface: Late Insertion and Spell Out in L2 Morphophonology*. New Sounds. Aarhus: University of Aarhus.

Backus, A. (1992). *Patterns of Language Mixing: A Study in Turkish-Dutch bilingualism*. Vol. 11. Otto Harrassowitz Verlag.

Badiola, L., Delgado, R., Sande, S., and Stefanich, S. (2017). Code-switching attitudes and their effects on acceptability judgement tasks. *Linguist. Approach. Biling.* 8, 5–24. doi: 10.1075/lab.16006.bad

Bandi-Rao, S., and den Dikken, M. (2014). "Light switches: On V as a pivot in codeswitching, and the nature of ban on word-internal switches," in *Grammatical Theory and Bilingual Codeswitching*, ed J. MacSwan (Cambridge, MA: MIT Press), 161–183.

Bartlett, L., and González-Vilbazo, K. E. (2013). The structure of the Taiwanese DP in Taiwanese–Spanish bilinguals: evidence from code-switching. *J. East Asian Ling.* 22, 65–99. doi: 10.1007/s10831-012-9098-3

Bessett, R. (2017). *The Integration of lone English nouns into bilingual Sonoran Spanish* (Unpublished doctoral dissertation). The University of Arizona, Tucson, AZ, United States.

Blanco-Elorrieta, E., and Pylkkanen, L. (2015). Brain bases of language selection: MEG evidence from Arabic-English bilingual language production. *Front. Hum. Neurosci.* 9:27. doi: 10.3389/fnhum.2015.00027

Blanco-Elorrieta, E., and Pylkkanen, L. (2016). Bilingual language control in perception versus action: MEG reveals comprehension control mechanisms in anterior cingulate cortex and domain-general control of production in dorsolateral prefrontal cortex. *J. Neurosci.* 36, 290–301. doi: 10.1523/JNEUROSCI.2597-15.2016

Bokamba, E. G. (1989). Are there syntactic constraints on code-mixing? *World Englishes* 8, 277–292. doi: 10.1111/j.1467-971X.1989.tb00669.x

Bolonyai, A. (2005). "English verbs in Hungarian/English code-switching," in *Proceedings of the 4th International Symposium on Bilingualism*, eds J. Cohen, K. McAlister, K. Rolstad, and J. MacSwan (Somerville, MA: Cascadilla Press), 317–327.

Borer, H. (2013). "The category of roots," in *The Roots of Syntax, the Syntax of Roots*, eds A. Alexiadou, H. Borer and F. Schaffer (Oxford: Oxford University Press), 112–148.

Budzhak-Jones, S. (1998). Against word-internal codeswitching: evidence from Ukrainian-English bilingualism. *Int. J. Biling.* 2, 161–182. doi: 10.1177/136700699800200203

Burkholder, M., Mathieu, E., and Sabourin, L. (2017). "The role of gender in mixed language nominal phrases: insights from distributed morphology," in *GALSA 14* (Southampton).

Cantone, K. (2003). "Evidence against a third grammar: code-switching in Italian-German bilingual children," in *Proceedings of the 4th International Sympoisum on Bilingualism*, eds J. Cohen, K. McAlister, K. Rolstad, K., and J. MacSwan (Somerville, MA: Cascadilla Press), 477–496.

Costa, A., and Sebastián-Gallés, N. (2014). How does the bilingual experience sculpt the brain? *Nat. Rev. Neurosci.* 15, 336–345. doi: 10.1038/nrn3709

Costa, R. S. (2009). Anglicismos léxicos en dos corpus. *Espéculo Rev Estud Liter.* 114. Available online at: https://www.biblioteca.org.ar/libros/150590.pdf

Damasio, A. (2018). *The Strange Order of Things: Life, Feeling and the Making of Cultures*. New York, NY: Pantheon.

De Belder, M., and Van Craenenbroeck, J. (2015). How to merge a root. *Linguist. Inq.* 46, 625–655. doi: 10.1162/LING_a_00196

Ebert, S., and Koronkiewicz, B. (2018). Monolingual stimuli as a foundation for analyzing code-switching data. *Linguist. Approach. Biling.* 8, 25–66. doi: 10.1075/lab.16007.ebe

Embick, D. (2015). *The Morpheme: A Theoretical Introduction*. Berlin: de Gruyter Mouton.

Embick, D., and Noyer, R. (2007). "Distributed morphology and the syntax/morphology interface," in *The Oxford Handbook of Linguistic Interfaces*, eds G. Ramchand and C. Reiss (Oxford: Oxford University Press), 289–324. doi: 10.1093/oxfordhb/9780199247455.013.0010

Fodor, J. A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. MIT Press.

Fuller, J. (1999). The role of English in Pennsylvania German Development: best supporting actress? *Am. Speech* 74, 36–55.

Garbin, G., Costa, A., Sanjuan, A., Forn, C., Rodriguez-Pujadas, A., Ventura, N., et al. (2011). Neural bases of language switching in high and early proficient bilinguals. *Brain Lang.* 119, 129–135. doi: 10.1016/j.bandl.2011.03.011

González-Vilbazo, K., Bartlett, L., Downey, S., Ebert, S., Heil, J., Hoot, B., et al. (2013). Methodological considerations in code-switching research. *Stud. Hispanic Lusophone Linguist.* 6, 119–138. doi: 10.1515/shll-2013-1143

González-Vilbazo, K., and Koronkiewicz, B. (2016). Tú y yo can codeswitch, nosotros cannot. *Spanish English Codeswitching Caribbean US* 11:237. doi: 10.1075/ihll.11.10gon

González-Vilbazo, K., and López, L. (2011). Some properties of light verbs in code-switching. *Lingua* 121, 832–850. doi: 10.1016/j.lingua.2010.11.011

González-Vilbazo, K. E. (2005). *Die syntax des code-switching. Esplugisch: sprachwechsel an der deutschen schule Barcelona* (Unpublished doctoral dissertation), Universität zu Köln, Köln, Germany.

Green, D. (2018). Language control and code-switching. *Languages* 3:8. doi: 10.3390/languages3020008

Green, D., and Kroll, J. (2019). "The neurolinguistics of bilingualism: plasticity and control," in *The Oxford Handbook of Neurolinguistics*, eds G. de Zubicaray and N. Schiller (Oxford: Oxford University Press), 261–294.

Green, D., and Wei, L. (2016). Code-switching and language control. *Biling. Lang. Cogn.* 19, 883–884. doi: 10.1017/S1366728916000018

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Biling. Lang. Cogn.* 1, 67–81. doi: 10.1017/S1366728998000133

Green, D. W., and Abutalebi, J. (2013). Language control in bilinguals: the adaptive control hypothesis. *J. Cogn. Psychol.* 25, 1–16. doi: 10.1080/20445911.2013.796377

Grimstad, M. B. (2017). The code-switching/borrowing debate: evidence from English-origin verbs in American Norwegian. *Ling. Linguaggio* XVI, 3–34. doi: 10.1418/86999

Grimstad, M. B., Lohndal, T., and Åfarli, T. A. (2014). Language mixing and exoskeletal theory: a case study of word-internal mixing in American Norwegian. *Nordlyd* 41, 213–237. doi: 10.7557/12.3413

Grosjean, F. (2001). "The bilingual's language modes," in *One Mind, Two Languages: Bilingual Language Processing*, ed J. Nicol (Oxford: Wiley-Blackwell), 122.

Halle, M. (2000). Distributed morphology: impoverishment and fission. *Amst. Stud. Theory Hist. Linguist. Sci. Series* 4, 125–150. doi: 10.1075/cilt.202.07hal

Halle, M., and Marantz, A. (1994). Some key features of distributed morphology. *MIT Working Papers Linguist.* 21:88.

Halmari, H. (1997). *Government and Codeswitching: Explaining American Finnish*. Vol. 12. Amsterdam: John Benjamins Publishing.

Harley, H. (2014). On the identity of roots. *Theor. Linguist.* 40, 225–276. doi: 10.1515/tl-2014-0010

Haspelmath, M. (2009). "Lexical borrowing: concepts and issues," in *Loanwords in the World's Languages: A Comparative Handbook*, eds M. Haspelmath and U. Tadmor (Berlin: De Gruyter Mouton), 35–54.

Haugen, J. D., and Siddiqui, D. (2013). Roots and the derivation. *Linguist. Inq.* 44, 493–517. doi: 10.1162/LING_a_00136

Hlavač, J. (1999). Phonological integration of English transfers in Croatian: evidence from the Croatian speech of second-generation Croatian-Australians. *Filologija* 32, 39–74.

Hlavac, J. (2000). *Croatian in Melbourne: Lexicon, switching and morphosyntactic features in the speech of second-generation bilinguals* (Unpublished doctoral dissertation), Monash University, Clayton, VIC, Australia.

Jackendoff, R. (1997). *The Architecture of the Language Faculty*. No. 28. Cambridge, MA: MIT Press.

Jackendoff, R. (2003). Précis of foundations of language : brain, meaning, grammar, evolution. *Behav. Brain Sci.* 26, 651–707. doi: 10.1017/S0140525X03000153

Jake, J. L., Myers-Scotton, C., and Gross, S. (2002). Making a minimalist approach to codeswitching work: adding the matrix language. *Biling. Lang. Cogn.* 5, 69–91. doi: 10.1017/S1366728902000147

Johannessenn, J. B. (2015). "The Corpus of American Norwegian Speech (CANS)," in *Proceedings of the 20th Nordic Conference of Computational Linguistics*, NEALT Proceedings Series 23, NODALIDA 2015, ed Béata Megyesi (Vilnius).

Johnson-Laird, P. N. (1980). Mental models in cognitive science. *Cogn. Sci.* 4, 71–115. doi: 10.1207/s15516709cog0401_4

Lohndal, T. (2013). Generative grammar and language mixing. *Theor. Linguist.* 39, 215–224. doi: 10.1515/tl-2013-0013

López, L., Alexiadou, A., and Veenstra, T. (2017). *Code Switching by Phase*. Chicago, IL: The University of Illinois at Chicago.

MacSwan, J. (1999). *A Minimalist Approach to Intrasentential Code Switching*. New York, NY: Garland Press.

MacSwan, J. (2000). The architecture of the bilingual language faculty: evidence from intrasentential code switching. *Biling. Lang. Cogn.* 3, 37–54. doi: 10.1017/S1366728900000122

MacSwan, J. (2005). Codeswitching and generative grammar: A critique of the MLF model and some remarks on "modified minimalism". *Biling. Lang. Cogn.* 8, 1–22. doi: 10.1017/S1366728904002068

MacSwan, J. (2014). *A Minimalist Approach to Intrasentential Code Switching*. Routledge.

MacSwan, J., and Colina, S. (2014). "Some consequences of language design: codeswitching and the PF interface," in *Grammatical Theory and Bilingual Codeswitching*, ed J. MacSwan (Cambridge, MA: MIT Press), 185–210.

Marantz, A. (1997). No escape from syntax: Don't try morphological analysis in the privacy of your own lexicon. *Univ. Pennsylvania Working Papers Linguist.* 4:14.

McGinnis, M. (2016). "Distributed morphology," in *The Cambridge Handbook of Morphology*, eds A. Hippisley and G. Stump (Cambridge: Cambridge University Press), 390–423.

Morrison, G. S. (2008). L1-Spanish Speakers' Acquisition of the English/i/—/I/Contrast: duration-based perception is not the initial developmental stage. *Lang. Speech* 51, 285–315. doi: 10.1177/0023830908099067

Myers-Scotton, C. (1992). Comparing codeswitching and borrowing. *J. Multiling. Multicult. Dev.* 13, 19–39. doi: 10.1080/01434632.1992.9994481

Myers-Scotton, C. (2000). Explaining the role of norms and rationality in codeswitching. *J. Pragmat.* 32, 1259–1271. doi: 10.1016/S0378-2166(99)00099-5

Poplack, S. (1980). Sometimes I'll start a sentence in Spanish y termino en español: toward a typology of codeswitching. *Linguistics* 18, 581–618. doi: 10.1515/ling.1980.18.7-8.581

Poplack, S. (1988). Contrasting patterns of code-switching in two communities. *Codeswitching Anthropol. Sociolinguist. Perspect.* 215:44.

Poplack, S., and Dion, N. (2012). Myths and facts about loanword development. *Lang. Var. Change* 24, 279–315. doi: 10.1017/S095439451200018X

Poplack, S., Sankoff, D., and Miller, C. (1988). The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics* 26, 47–104. doi: 10.1515/ling.1988.26.1.47

Riksem, B., Grimstad, M., Lohndal, T., and Afarli, T. (2019). Language mixing within verbs and nouns in American Norwegian. *J. Comp. Germanic Syntax* 22, 189–209. doi: 10.1007/s10828-019-09109-6

Rumelhart, D. E., McClelland, J. L., and PDP Research Group. (1987). *Parallel Distributed Processing*. Vol. 1. Cambridge, MA: MIT Press.

Schütze, C. T., and Sprouse, J. (2014). "Judgment data," in *Research Methods Linguist,* 27–50. doi: 10.1017/CBO9781139013734.004

Sharwood Smith, M., and Truscott, J. (2014). *The Multilingual Mind: A Modular Processing Perspective*. Cambridge: Cambridge University Press.

Stammers, J., and Deuchar, M. (2012). Testing the nonce borrowing hypothesis: counter-evidence from English-origin verbs in Welsh. *Biling. Lang. Cogn.* 15, 630–643. doi: 10.1017/S1366728911000381

Stefanich, S. (2019). *A morphophonological account of Spanish/English word-internal codeswitching* (Unpublished doctoral dissertation). University of Illinois at Chicago, Chicago, IL, United States.

Stefanich, S., and Cabrelli Amaro, J. (2018a). "Phonological factors of Spanish/English word-internal code-switching," in *Code-Switching: Theoretical Questions, Experimental Answers*, ed L. López (Amsterdam: John Benjamins), 195–222.

Stefanich, S., and Cabrelli Amaro, J. (2018b). "Spanish/English word-internal code-switching: the case of /ɪ/ and /θ/," in *Paper Presented at the Hispanic Linguistics Symposium* (Austin, TX).

Treffers-Daller, J. (1993). *Mixing Two Languages: French-Dutch Contact in a Comparative Perspective*. Berlin: Mouton de Gruyter.

Truscott, J., and Sharwood Smith, M. (2004). Acquisition by processing: a modular perspective on language development. *Biling. Lang. Cogn.* 7, 1–20. doi: 10.1017/S1366728904001178

Truscott, J., and Sharwood Smith, M. (2016). *Representation, Processing and Code-switching*. Bilingualism: Language and Cognition, 1–14.

Turker, E. (2000). *Turkish-Norwegian codeswitching: evidence from intermediate and second- generation Turkish immigrants in Norway* (Un-published PhD dissertation) University of Oslo, Oslo, Norway.

Van Dijk, T. A. (1997). "Cognitive context models and discourse," in *Language Structure, Discourse and The Access to Consciousness*, ed M. Stamenov (Amsterdam: John Benjamins), 189–226.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Communication

ORIGINAL RESEARCH
published: 08 April 2020
doi: 10.3389/fcomm.2020.00017

Check for updates

Modeling Morphological Priming in German With Naive Discriminative Learning

R. Harald Baayen[1]* *and Eva Smolka*[2]

[1] Department of Linguistics, University of Tübingen, Tübingen, Germany, [2] Department of Linguistics, University of Konstanz, Konstanz, Germany

Both localist and connectionist models, based on experimental results obtained for English and French, assume that the degree of semantic compositionality of a morphologically complex word is reflected in how it is processed. Since priming experiments using English and French morphologically related prime-target pairs reveal stronger priming when complex words are semantically transparent (e.g., *refill–fill*) compared to semantically more opaque pairs (e.g., *restrain–strain*), localist models set up connections between complex words and their stems only for semantically transparent pairs. Connectionist models have argued that the effect of transparency should arise as an epiphenomenon in PDP networks. However, for German, a series of studies has revealed equivalent priming for both transparent and opaque prime-target pairs, which suggests mediation of lexical access by the stem, independent of degrees of semantic compositionality. This study reports a priming experiment that replicates equivalent priming for transparent and opaque pairs. We show that these behavioral results can be straightforwardly modeled by a computational implementation of Word and Paradigm Morphology (WPM), Naive Discriminative Learning (NDL). Just as WPM, NDL eschews the theoretical construct of the morpheme. NDL succeeds in modeling the German priming data by inspecting the extent to which a discrimination network pre-activates the target lexome from the orthographic properties of the prime. Measures derived from an NDL network, complemented with a semantic similarity measure derived from distributional semantics, predict lexical decision latencies with somewhat improved precision compared to classical measures, such as word frequency, prime type, and human association ratings. We discuss both the methodological implications of our results, as well as their implications for models of the mental lexicon.

Keywords: morphological processing, naive discriminative learning, priming, semantic transparency, stem-based lexical access, complex verbs, morphological priming

OPEN ACCESS

Edited by:
Gary Libben,
Brock University, Canada

Reviewed by:
Pilar Ferré Romeu,
University of Rovira i Virgili, Spain
Helena Mendes Oliveira,
University of Minho, Portugal

Correspondence:
R. Harald Baayen
harald.baayen@uni-tuebingen.de

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *08 October 2019*
Accepted: *10 March 2020*
Published: *08 April 2020*

Citation:
Baayen RH and Smolka E (2020) Modeling Morphological Priming in German With Naive Discriminative Learning. Front. Commun. 5:17. doi: 10.3389/fcomm.2020.00017

1. INTRODUCTION

Current mainstream models of lexical processing assume that complex words such as *unmanagability* comprise several morphemic constituents, *un-*, *manage*, *-able*, and *-ity*, that recur in the language in many other words. Since early research in the seventies (e.g., Taft and Forster, 1975), it has been argued that the recognition of morphologically complex words is mediated by such morphemic units (for a review of models of morphological processing, see Milin et al., 2017b).

One of the issues under investigation in this line of research is whether visual input is automatically decomposed into morphemes before semantics is accessed. Several studies have argued in favor of early morpho-orthographic decomposition (Longtin et al., 2003; Rastle et al., 2004; Rastle and Davis, 2008), but others argue that semantics is involved from the start (Feldman et al., 2009), that the effect is task dependent and is limited to the lexical decision task (Norris and Kinoshita, 2008; Dunabeitia et al., 2011; Marelli et al., 2013), or fail to replicate experimental results central to decompositional accounts (Milin et al., 2017a).

Another issue that is still unresolved is whether complex words are potentially accessed through two routes operating in parallel, one involving decomposition and the other whole-form based retrieval (Frauenfelder and Schreuder, 1992; Marslen-Wilson et al., 1994b; Baayen et al., 1997, 2003). Recent investigations that make use of survival analysis actually suggest that whole-word based effects precede in time constituent-based effects (Schmidtke et al., 2017).

A third issue concerns the role of semantic transparency. Priming studies conducted on English and French prefixed derivations that are semantically transparent, such as *distrust*, have reported facilitation of the recognition of their stems (*trust*), as well as other prefixed or suffixed derivations, such as *entrust* or *trustful*. The same holds for suffixed derivations that are semantically transparent like *production* and *productivité* in French or *confession* and *confessor* in English, which prime each other and their stem (*confess*). The critical condition in this discussion, however, concerns semantically opaque (i.e., non-compositional) derivations, such as *successor*, which appear not to facilitate the recognition of stems like *success*. This latter finding was replicated under auditory prime presentations or visual priming at long exposure durations at 230 or 250 ms (e.g., Rastle et al., 2000; Feldman and Prostko, 2001; Pastizzo and Feldman, 2002; Feldman et al., 2004; Meunier and Longtin, 2007; Lavric et al., 2011). Localist accounts take these findings to indicate that only semantically transparent complex words are processed decompositionally, via their stem, while semantically opaque words are processed as whole word units. Although on different grounds, also distributed connectionist approaches assume that the facilitation between complex words and their stem depends on their meaning relation. In a series of cross-modal priming experiments, Gonnerman et al. (2007) showed for English that morphological effects vary according to the gradual overlap of form and meaning between word pairs. Indeed, word pairs with a strong phonological and semantic relation like *preheat–heat* induced stronger priming than words with a moderate phonological and semantic relation like *midstream–stream*, and words holding a low semantic relation like *rehearse–hearse* induced no priming at all. According to connectionist accounts of lexical processing, this result arises as the consequence of the extent to which orthographic, phonological, and semantic codes converge.

However, these findings for English and French contrast with results repeatedly obtained for German, where morphological priming appears to be unaffected by semantic transparency (Smolka et al., 2009, 2014, 2015, 2019). Under auditory or overt visual prime presentations, morphologically related complex verbs facilitated the recognition of their stem regardless of whether they were semantically transparent (*aufstehen–stehen*, "stand up"–"stand") or opaque (*verstehen–stehen*, "understand"–"stand"). Smolka et al. interpreted these findings to indicate that a German native speaker processes a complex verb like *verstehen* by accessing the stem *stehen* irrespective of the whole-word meaning, and argued that morphological structure overrides meaning in the lexical processing of German complex words. To account for such stem effects without effects of semantic transparency, they hypothesized a model in which the frequency of the stem is the critical factor, such that stems of complex words are accessed and activated, independent of the meaning composition of the complex word.

These findings for German receive support from experiments on Dutch—a closely related language with a highly similar system of verbal prefixes, separable particles, and non-separable particles. Work by Schreuder et al. (1990), using an intramodal visual short SOA partial priming technique to study Dutch particle verbs, revealed morphological effects without modulation by semantic transparency. Experiments addressing speech production in Dutch (Roelofs, 1997a,b; Roelofs et al., 2002) likewise observed, using the implicit priming task, that priming effects were equivalent for transparent and opaque prime-target pairs. Morphological priming without effects of semantic transparency have recently been replicated in Dutch under overt prime presentations (Creemers et al., 2019; De Grauwe et al., 2019). Unprimed and primed visual lexical decision experiments on Dutch low-frequency suffixed words with high-frequency base words revealed that the semantics of opaque complex words were equally quickly available as the semantics of transparent complex words (Schreuder et al., 2003), contradicting the original prediction of this study that transparent words would show a processing advantage compared to their opaque counterparts.

Importantly, there are some studies in English, e.g., Gonnerman et al. (2007, Exp. 4) and Marslen-Wilson et al. (1994a, Exp. 5), that applied a similar cross-modal priming paradigm with auditory primes and visual targets, and with similar prefixed stimuli as in the abovementioned studies by Smolka and collaborators, but found no priming for semantically opaque pairs like *rehearse–hearse* (for similar ERP-results in English see Kielar and Joanisse, 2011). Thus, results for German and results for English appear at present to be genuinely irreconcilable[1].

In what follows, we first present an overt visual priming experiment that provides further evidence for the equivalent facilitation effects seen for German transparent and opaque prime-target pairs. The behavioral results are consistent with localist models in which connections between stems and derived words are hand-wired into a network, as argued by, e.g., Smolka et al. (2007, 2009, 2014, 2015) and Smolka and Eulitz (2018). However, this localist model is a *post-hoc* description of the experimental findings, and a computational implementation for this high-level theory is not available.

[1]See Smolka et al. (2014) and Günther et al. (2019) for possible explanations.

In this study, we proceed to show that the observed stem priming effects can be straightforwardly modeled by naive discriminative learning (NDL, Baayen et al., 2011, 2016a,b; Arnold et al., 2017; Divjak et al., 2017; Sering et al., 2018b; Tomaschek et al., 2019) without reference to stems or other morphological units, and without requiring hand-crafting of connections between such units. In fact, measures derived from an NDL network, complemented with a semantic similarity measure derived from distributional semantics, turn out to predict lexical decision latencies with greater precision compared to classical measures, such as word frequency, prime type, and semantic association ratings. Importantly, the NDL model predicts the effects of stem priming without a concomitant effect of semantic compositionality. According to the NDL model, the crucial predictor is the extent to which the target is pre-activated by the sublexical form features of the prime. In the final section, we discuss both the methodological implications of our results, as well as their implications for models of the mental lexicon.

2. THE PRIMING EXPERIMENT

2.1. Previous Results for German Complex Verbs

German complex verbs present a very useful means to study the effects of morphological structure with or without meaning relatedness to the same base verb. German complex verbs are very productive and frequently used in standard German. The linguistic literature (Fleischer and Barz, 1992; Eisenberg, 2004) distinguishes two word formations: prefix verbs and particle verbs. Both consist of a verbal root and either a verbal prefix or a particle.

In spite of some prosodic and morphosyntactic differences (see Smolka et al., 2019), prefix and particle verbs share many similar semantic properties. Both may differ in the degree of semantic transparency with respect to the meaning of their base. For example, the particle *an* ("at") only slightly alters the meaning of the base *führen* ("guide") in the derivation *anführen* ("lead"), but radically does so with respect to the base *schicken* ("send") in the opaque derivation *anschicken* ("get ready"). Similarly, the prefix *ver-* produces the transparent derivation *verschicken* ("mail") as well as the opaque derivation *verführen* ("seduce"). Prefix and particle verbs are thus a particularly useful means by which the effects of meaning relatedness to the same base verb can be studied. For instance, derivations of the base *tragen* ("carry"), such as *hintragen* ("carry to"), *forttragen* ("carry away"), *zurücktragen* ("carry back"), *abtragen* ("carry off"), *auftragen* ("apply"), *vertragen* ("get along"), *ertragen* ("suffer"), alter the meaning relatedness from fully transparent to fully opaque with respect to the base. It is important to note that, in general, complex verbs in German are true etymological derivations of their base, regardless of the degree of semantic transparency they share with it. Because morphological effects of prefix and particle verbs are alike in German (see Smolka and Eulitz, 2018; Smolka et al., 2019) and Dutch (Schriefers et al., 1991), henceforth, we refer to them as "complex verbs" or "derived verbs."

Previous findings on complex verbs in German have shown that these verbs strongly facilitate the recognition of their stem, without any effect of semantic transparency (Smolka et al., 2009, 2014, 2015, 2019; Smolka and Eulitz, 2018). That is, semantically opaque verbs, such as *verstehen* ("understand") primed their base *stehen* ("stand") to the same extent as did transparent verbs, such as *aufstehen* ("stand up"). Further, the priming by both types of morphological primes was stronger than that by either purely semantically related primes like *aufspringen* ("jump up") or purely form-related primes like *bestehlen* ("steal"). The morphological effects remained unaffected by semantic transparency under conditions that were sensitive to detecting semantic and form similarity, that is, when semantic controls like *verlangen–fordern* ("require"–"demand") and *Biene–Honig* ("bee"–"honey") induced semantic facilitation or when form-controls with embedded stems, as in *bekleiden–leiden* ("dress"–"suffer") and *Bordell–Bord* ("brothel"–"board") induced form inhibition (see Exp. 3 in Smolka et al., 2014). This offered assurance that the lack of a semantic transparency effect between semantically transparent and opaque complex verbs was not a null effect but rather indicated that morphological relatedness overrides both semantic and form relatedness.

Further studies explored the circumstances of stem facilitation in more detail. For example, in spite of several differences in the phonological and morpho-syntactic properties of prefix and particle verbs, prefix verbs showed processing patterns that were substantially the same as those for particle verbs and, crucially, were uninfluenced by semantic transparency (Smolka et al., 2019). Furthermore, stem access occurs regardless of the directionality of prime and target *entwerfen–werfen* vs. *werfen–entwerfen* vs. *entwerfen–bewerfen* (Smolka and Eulitz, 2011).

Stem access is modality independent, as it occurs under both intra-modal (visual-visual) and cross-modal (auditory-visual) priming conditions (Smolka et al., 2014, 2019). Finally, event-related brain potentials revealed wide-spread N400 brain potentials in response to semantically transparent and opaque verbs without effects of semantic transparency—N400 brain potentials that are generally taken to be characteristic to indicate expectancy and (semantic) meaning integration. Most importantly, these brain potentials revealed that stem facilitation in German occurs without an overt behavioral response and is stronger than the activation by purely semantically related verbs or form-related verbs (Smolka et al., 2015).

The present experiment was closely modeled after previous experiments addressing priming effects for German verb pairs (e.g., Smolka et al., 2009, 2014).

2.2. Design

We compared the differential effects of semantic, form, or morphological relatedness between complex verbs and a base verb in four priming conditions: (a) semantic condition, where the complex verb was a synonym of the target verb, (b) morphological transparent condition, where the complex verb was a semantically transparent derivation of the target verb, (c) morphological opaque condition, where the complex verb was a semantically opaque derivation of the target verb, and (d) form condition, where the base of the complex verb was form-related

TABLE 1 | Stimulus characteristics of related primes and their matched unrelated controls in the semantic synonym list, semantically transparent list, semantically opaque list, and form control list.

List	Relatedness	Lemma frequency	Word length	Syllable length	Age of acquisition	Relatedness score
SEMANTIC						
vorangehen–führen	Related	4.9	9.2	3.0	5.4	5.5
("antecede"–"guide")		(6.3; 0–24)	(1.2; 7–12)	(0.2; 3–4)	(1.3; 3.4–7.4)	(0.6; 4.3–6.6)
	Unrelated	4.7	9.1	3.0	7.1	
		(6.1; 0–24)	(1.1; 7–11)	(0; 3–3)	(1.7; 4.3–9.8)	
TRANSPARENT						
anführen–führen	Related	5.5	9.3	3.0	5.2	5.7
("head"–"guide")		(5.6; 0–17)	(1.5; 7–11)	(0; 3–3)	(1.2; 3.1–8.3)	(0.5; 4.8–6.6)
	Unrelated	5.6	9.2	3.0	6.5	
		(5.8; 0–20)	(1.1; 7–11)	(0; 3–3)	(1.7; 3.7–9.2)	
OPAQUE						
verführen–führen	Related	8.8	9.6	3.1	8.0	2.1
("seduce"–"guide")		(9.5; 0–34)	(1.4; 8–13)	(0.3; 3–4)	(1.7; 4.4–9.8)	(0.5; 1.1–2.9)
	Unrelated	9.1	9.1	3.0	6.6	
		(10.7; 0–47)	(1.1; 7–11)	(0.2; 3–4)	(1.6; 3.7–9.2)	
FORM						
befühlen–führen	Related	5.8	9.7	3.0	7.7	1.3
("palpate"–"guide")		(9.2; 0–36)	(1.2; 8–12)	(0; 3–3)	(1.8; 4.7–10.3)	(0.5; 1–2.9)
	Unrelated	6.4	9.2	3.0	6.5	
		(10.9; 0–47)	(1.2; 7–11)	(0.2; 3–4)	(1.6; 3.4–9.2)	

Statistics are given for the total set of stimuli: mean (SD; range); sample stimuli of prime-target pairs are italicized. Frequencies are from the CELEX database (Baayen et al., 1993), count is per million. Lists were between-subjects.

with the base of the target. We measured lexical decision latencies to the target verbs and calculated priming relative to an unrelated control condition. In addition to the (unrelated minus related) priming effects, the influence of the stem should surface in the comparison of the conditions (a) and (b), where both types of primes are synonyms of the base verb—the former holding a different stem as the target, the latter holding the same stem as the target; the influence of the degree of semantic transparency should surface in the comparison between conditions (b) and (c), where both types of primes are true morphological derivations of the base target. The influence of form similarity should surface in the comparison between conditions (c) and (d), where both types of primes have stems that are form-similar with the target.

As in our previous experiments, we were interested in tapping into lexical processing, when participants are aware of the prime and integrate its meaning, and thus applied overt visual priming at a long SOA (see Milin et al., 2017b, for a comparison between masked and overt priming paradigms). We used only verbs as materials to avoid word category effects, and inserted a large number of fillers to prevent expectancy or strategic effects. Different from our previous experiments, though, we applied a between-subject and between-target design.

In summary, the primes in all conditions were complex verbs with the same morphological structure and were thus (a) of the same word category, and (b) closely matched on distributional variables like lemma frequency, number of syllables and letters. They differed only with respect to the morphological, semantic, or form-relatedness with the target. Prime conditions are exemplified in **Table 1**; all critical items are listed in **Appendix A**. Our prediction is that both semantically transparent and opaque complex verbs will induce the same amount of priming to their base, and that this priming will be stronger than the priming by either semantically related or form-related verbs.

2.3. Method

2.3.1. Participants

Fifty students of the University of Konstanz participated in the experiment (14 males; mean age = 22.69, range 19–32). All were native speakers of German, were not dyslexic, and had normal or corrected-to-normal vision. They were paid for their participation.

2.3.2. Materials

As critical stimuli, 88 prime-target pairs with complex verbs as primes and base verbs as targets were selected from the CELEX German lexical database (Baayen et al., 1993), 22 pairs in each of four conditions (see also **Table 1**): (a) morphologically unrelated synonyms of the base (e.g., *vorangehen–führen*, "antecede"–"guide"), (b) morphologically related synonyms of the base, these were semantically transparent derivations of the base, (e.g., *anführen–führen*, "head"–"guide"), (c) semantically opaque derivations of the base (e.g., *verführen–führen*, "seduce"–"guide"), and (d) semantically and morphologically unrelated form controls that changed a letter in the stem (by retaining the stem's onset and changing a letter in the rime, e.g., *befühlen–führen*, "palpate"–"guide"). Complex verbs in conditions (a) and

(b) were synonyms of the target base and were selected by means of the online synonym dictionaries http://www.canoo.net/ and https://synonyme.woxikon.de/.

For each of the 88 related primes, we selected an unrelated control that served as baseline and (a) was morphologically, semantically, and orthographically unrelated to the target and (b) matched the related prime in word class, morphological complexity (i.e., it was a complex verb), number of letters and syllables. In addition, control primes were pair-wise matched to the related primes on lemma frequency according to CELEX. Furthermore, primes across conditions were matched on lemma frequency according to CELEX.

The critical set of 88 prime-target pairs was selected from a pool of verb pairs that had been subjected to semantic association tests, in which participants rated the meaning relatedness between the verbs of each prime-target pair on a 7-point scale from completely unrelated (1) to highly related (7) (for a detailed description of the database see Smolka and Eulitz (2018). The following criteria determined whether a verb pair was included in the critical set: The mean ratings for a semantically-related pair (in the synonym and semantically transparent conditions) had to be higher than 4, and those for a semantically unrelated pair (in the semantically opaque and form-related conditions) lower than 3. The set of words that were included in the experiment had mean ratings of 5.5 (range 4.3–6.7) for synonyms, 5.7 (range 4.78–6.56) for semantically transparent derivations, 2.13 (range 1.5–2.88) for semantically opaque derivations, and 1.7 (range 1.0–2.89) for form-related pairs. **Table 1** provides the prime characteristics (lemma frequency, number of letters and syllables, meaning relatedness); the **Appendix** lists all stimuli.

In order to prevent strategical processes, a total of 140 prime-target pairs was added as fillers. All had complex verbs as primes, 48 had verbs and 92 had pseudoverbs as targets. With respect to the former, 18 of the 48 prime-verb fillers comprised related prime-target pairs of the other lists. These types were included to assure that participants would not detect a certain type of prime-target relatedness in a list. For example, list A held six items of list B, six of list C, and six of list D as fillers. The other 30 prime-target pairs were semantically, morphologically, and orthographically unrelated.

Regarding the prime-pseudoverb fillers, 44 of the pseudoverb targets were closely matched to the critical verb targets by keeping the onset of the verbs' first syllable (e.g., *binden–binken*). To further ensure that participants did not respond with "word" decisions for any trial where prime and target were orthographically similar, eleven pseudoverbs were preceded by a form-related prime (e.g., *umwerben–wersen*) to mimick the form condition. All pseudoverbs were constructed by exchanging one or two letters in real verbs, while preserving the phonotactic constraints of German.

The between-subject design had the following list composition: Each list comprised 184 prime-target pairs, half of these holding verbs, the other half pseudoverbs as targets. Of the 92 prime-verb pairs, 22 were related prime-target pairs of either condition (a), (b), (c), or (d), 22 were matched unrelated prime-target pairs, and 48 were filler pairs (30 unrelated and 18 of other related conditions). The 92 prime-pseudoverb-target pairs included 44 form-matched and 48 unrelated pairs.

Overall, the large amount of fillers in the present study reduced the proportion of (a) critical prime-target pairs to 24% per list or 48% of prime-verb pairs, and (b) related prime-target pairs (including critical and filler pairs) to 22% per list or 43% of verb pairs. Napps and Fowler (1987) showed that a reduction in the proportion of related items from 75% to 25% reduced both facilitatory and inhibitory effects. A significant reduction of related items in the present study should thus discourage participants from expecting a particular related verb target and thus prevent both expectancy and failed expectancy effects. All filler items differed from the critical items. Throughout the experiment, all primes and targets were presented in the infinitive (stem/-en), which is also the citation form in German.

2.3.3. Apparatus

Stimuli were presented on a 18.1″ monitor, connected to an IBM-compatible AMD Atlon 1.4 GHz personal computer. Stimulus presentation and data collection were controlled by the Presentation software developed by Neurobehavioral Systems (https://www.neurobs.com). Response latencies were recorded from the left and right buttons of a push-button box.

2.3.4. Procedure

Each participant saw only one list. Each list was divided into four blocks, each block containing the same amount of stimuli per condition. The critical prime-target pairs were rotated over the four blocks according to a Latin Square design in such a way that the related and unrelated primes of the same target were separated by a block. The related fillers (form-related prime-pseudoverb pairs, related prime-verb pairs) and unrelated filler pairs were evenly allocated to the blocks.

In total, an experimental session comprised 184 prime-target pairs, with 66 pairs per block. Within blocks, prime-target pairs were randomized separately for each participant. Twenty additional prime-target pairs were used as practice trials. Participants were tested individually in a dimly lit room, seated at a viewing distance of about 60 cm from the screen. Stimuli were presented in Sans-Serif letters on a black background. To ensure that primes and targets were perceived as physically distinct stimuli, primes were presented in uppercase letters, point 32, in light blue (RGB: 0-255-255), 20 points above the center of the screen. Targets were presented centrally in lowercase letters, point 36, in yellow (RGB: 255-255-35).

Each trial started with a fixation cross in the center of the screen for 300 ms. This was followed by the presentation of the prime for 400 ms, followed by an offset (i.e., a blank screen) for 100 ms, resulting in a stimulus onset asynchrony (SOA) of 500 ms. After the offset, the target immediately followed and remained on the screen until a participant's response. The intertrial interval was 1,500 ms. Participants were instructed to make lexical decisions to the targets, as fast and as accurately as possible. "Word" responses were given with the index finger of the dominant hand, "pseudoword" responses with the subordinate hand. Feedback was given on both correct ("richtig") and incorrect ("falsch") responses during the practice session, and on incorrect responses during the experimental session. The experiment lasted for about 12 min, during which participants self-administered the breaks between blocks.

2.4. Results

A generalized additive mixed model (Wood, 2017) was fitted to the inverse-transformed reaction times with predictors Prime Type (using treatment coding, with the unrelated condition as reference level) and log target frequency[2]. Random intercepts were included for target and prime, and a factor smooth for the interaction of subject by trial number (see Baayen et al., 2017, for detailed discussion of this non-linear counterpart to what in a linear mixed model would be obtained with by-subject random intercepts and by-subject random slopes for trial). **Table 2** presents the model summary. Prime-target pairs in the semantic condition were responded to slightly more quickly than prime-target pairs in the unrelated condition. Prime-target pairs in the transparent and opaque conditions showed substantially larger facilitation of equal magnitude. Prime-target pairs in the form condition elicited reaction times that did not differ from those seen in the control condition.

To obtain further insight into the effects of the predictors not only for the median, but across the distribution of reaction times, we fitted quantile GAMs to the deciles 0.1, 0.2, ..., 0.9, using the **qgam** package (Fasiolo et al., 2017)[3]. For the median, the quantile GAM also complements the Gaussian GAMM reported in **Table 2**. The Gaussian GAMM could have been expanded with further random effects for the interaction of subject by priming effect, but such models run the risk of overspecification (Bates et al., 2015). More importantly, the distribution of the residuals showed clear deviation from normality that resisted correction. As quantile GAMS are distribution free, simple main effects can

TABLE 2 | Generalized additive mixed model fitted to inverse-transformed primed lexical decision latencies.

A. Parametric coefficients	Estimate	Std. Error	*t*-value	*p*-value
Intercept (PrimeType = Unrelated)	−1.8791	0.0298	−63.0083	<0.0001
PrimeType = Semantic	−0.0761	0.0240	−3.1709	0.0015
PrimeType = Transparent	−0.2514	0.0248	−10.1222	<0.0001
PrimeType = Opaque	−0.2519	0.0249	−10.1305	<0.0001
PrimeType = Form	0.0016	0.0242	0.0669	0.9467
B. Smooth terms	**edf**	**Ref.df**	***F*-value**	***p*-value**
TPRS smooth LogTargetFreq	1.0001	1.0001	9.1823	0.0025
Factor smooths for trial × subject	113.0046	449.0000	2.6135	<0.0001
Random intercepts Prime	16.7699	116.0000	0.1907	0.0585
Random intercepts Target	21.4123	39.0000	1.6359	<0.0001

be studied without having to bring complex random effects into the model as a safeguard against anti-conservative *p*-values.

Figure 1 presents, from top left to bottom right, the effects of Prime Type for the deciles 0.1, 0.2, ..., 0.9. The *p*-values above the bars concern the contrasts with the unrelated condition (the reference level). Across the distribution, the form condition was never significantly different from the unrelated condition. The small effect of the semantic condition hardly varied in magnitude across deciles, but was no longer significant at the last decile. The magnitude of the effects for the transparent and opaque conditions was significantly different from that for the unrelated condition across all deciles, and increased in especially the last three deciles. Across the deciles, the transparent condition showed a growing increase in facilitation compared to the opaque condition, but as indicated by the *p*-values in red, the difference between these two conditions was never significant.

Figure 2 visualizes the effect of target frequency. From the second decile onwards, target frequency was significant, with greater frequencies affording shorter reaction times, as expected. The magnitude of the frequency effect, as well as its confidence interval, increased across deciles.

In summary, replicating earlier studies, morphologically related primes elicited a substantial priming effect that did not vary with semantic transparency. The priming effect tended to be somewhat stronger at larger deciles, for which the effect of target frequency was also somewhat larger.

When the transparent and opaque priming conditions are considered in isolation, it might be argued that participants are superficially scanning primes and targets for shared stems, providing a word response when a match is found, and a non-word response otherwise. Such a task strategy can be ruled out, however, due to the high proportion of unrelated fillers, and once the other two priming conditions are taken into consideration. If subjects would indeed have been scanning for stems shared by prime and target, then the Unrelated and Semantic conditions should have elicited the same mean reaction times, contrary to fact. Clearly, participants processed the word stimuli more deeply than a form-based scan suggests. Furthermore, since

[2]Frequency was added as a covariate to the model for three reasons. First, frequency varied *within* priming conditions, and since frequency is a strong predictor of response latencies in the visual lexical decision task, the presence of a frequency effect certifies that participants were engaging with the task, rather than performing on the basis of some high-level task-specific strategies, such as searching for similar letter substrings. Second, we wanted to make sure that any differences in frequency between priming conditions, related not to mean frequencies, but to imbalances in the distributions of frequencies within the separate priming conditions, cannot be held responsible for the observed priming effects. Third, we were interested in exploring to what extent the magnitude of the frequency effect might change across the distribution of reaction times.

[3]Quantile regression is a statistical method for predicting the quantiles of a response variable. Standard regression methods consider only the mean of the response, which, if the response is Gaussian, is equal to the median, which in turn is the 5th decile. However, a researcher's interest may be not in predicting the mean, but a specific quantile. For instance, an electricity company may be interested in balancing atomic power and water power. Energy from water power is scarce, but the (limited) amount of power generated can be varied quickly in time as demand changes. Energy from atomic plants is available in large quantities, but the amount of energy generated can be varied only very slowly over time. To balance atomic power and water power, it can be of interest to generate 98% of the energy demand using atomic power, and supplement the remaining energy demand from water power. In this case, quantile regression can be used to predict the 98 percentile of energy demand as a function of time (Fasiolo et al., 2017). For the analysis of reaction times, quantile regression can be used to clarify where in the distribution of reaction times effects are present, and to investigate whether effects change in magnitude across the distribution (see e.g., Tomaschek et al., 2018, for an application in phonetics).

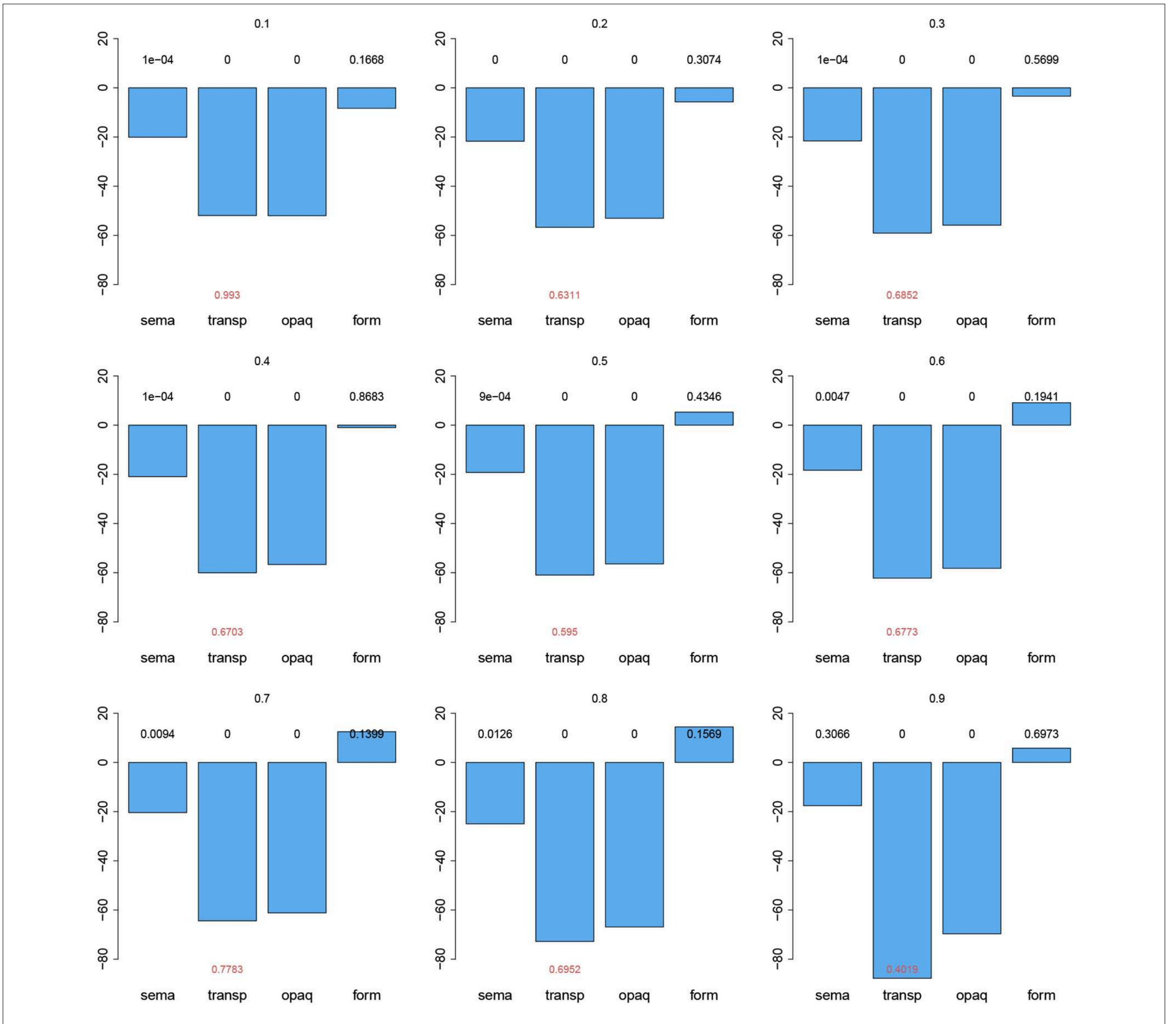


FIGURE 1 | Effects of Prime Type (with unrelated as reference level) in a Quantile GAM fitted to primed lexical decision latencies, for deciles 0.1, 0.2, ..., 0.9.

form-based evaluation of pairs in the Form condition is more difficult given the substantial similarity of the stems, longer reaction times are predicted for the Form condition compared to the Unrelated condition. Also this prediction is falsified by the data: The means for the Form and Unrelated condition do not differ significantly.

At first sight, it might be argued that the present between-participants design has strengthened the opaque priming effects. However, given that all our previously conducted experiments used within-participants designs (e.g., three visual priming experiments in Smolka et al., 2009; one cross-modal and two visual priming experiments in Smolka et al., 2014; two cross-modal priming experiments in Smolka et al., 2019; and one visual EEG priming experiment in Smolka et al., 2015) and yielded equivalent priming by opaque and transparent prime-target pairs, we are confident that our present findings are not due to a design limitation.

We therefore conclude that the behavioral results of "pure morphological priming" without semantic transparency effects in German, as well as the older results obtained for speech production in Dutch, appear to indicate a fundamental role in lexical processing for morphemic units, such as the stem. However, perhaps surprisingly, developments in current linguistic morphology indicate that the theoretical construct of the morpheme is in many ways problematic. In what follows, we show that the present results can be explained within the framework of naive discriminative learning, even though this theory eschews morphemic units altogether.

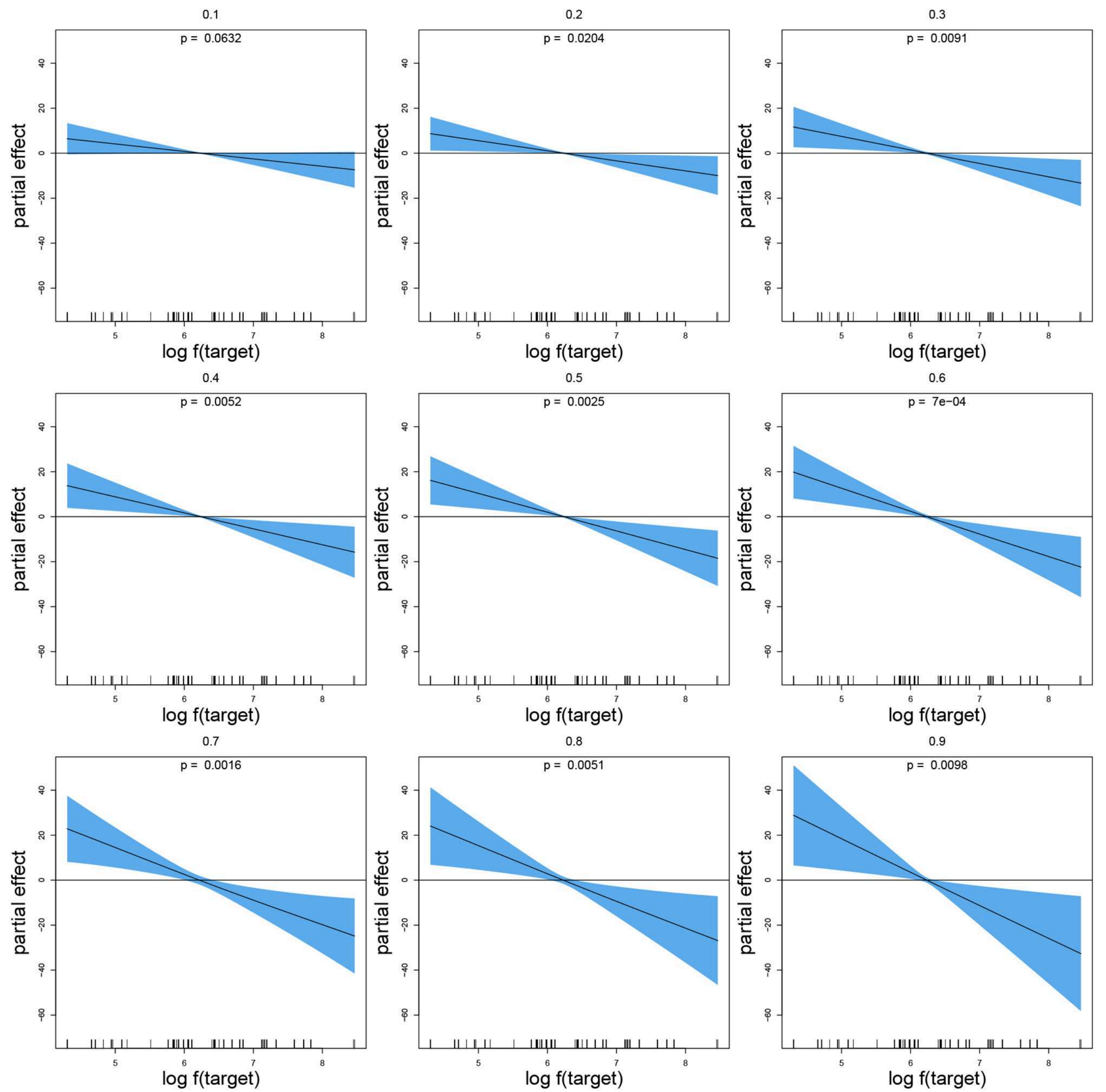


FIGURE 2 | Partial effect of log target frequency in a Quantile GAM fitted to primed lexical decision latencies, for deciles 0.1, 0.2, …, 0.9.

3. COMPUTATIONAL MODELING WITH NAIVE DISCRIMINATIVE LEARNING

Before introducing naive discriminative learning (NDL), we first provide a brief overview of developments in theoretical morphology over the last decades that motivated the development of NDL.

3.1. Developments in Theoretical Morphology

The concept of the morpheme, as the minimal linguistic sign combining form and meaning, traces its history to the American structuralists that sought to further systematize the work of Leonard Bloomfield (see Blevins, 2016, for detailed discussion). The morpheme as minimal sign has made it into many introductory textbooks (e.g., Plag, 2003; Butz and Kutter, 2016). The hypothesis that semantically transparent complex words are processed compositionally, whereas semantically opaque words are processed as units, is itself motivated by the belief that morphemes are linguistic signs. For semantically opaque words, the link between form and meaning is broken, the morpheme is no longer a true sign, and hence the rules operating over true signs in comprehension and production are no longer relevant.

The theoretical construct of the morpheme as smallest sign of the language system has met with substantial criticism (see e.g., Matthews, 1974; Beard, 1977; Aronoff, 1994; Stump, 2001; Blevins, 2016). Whereas the morpheme-as-sign appears a reasonably useful construct for agglutinating languages, such as Turkish, as well as for morphologically simple languages, such as English (but see Blevins, 2003), it fails to provide much insight for typologically very dissimilar languages, such as Latin, Estonian, or Navajo (see e.g., Baayen et al., 2018, 2019; Chuang et al., 2019, for detailed discussion). One important insight from theoretical morphology is that systematicities in form are not coupled in a straightforward one-to-one way with systematicities in meaning. Realizational theories of morphology (e.g., Stump, 2001) therefore focus on how sets of semantic features are expressed in phonological form, without seeking to find atomic form features that line up with atomic semantic features. Interestingly, as pointed out by Beard (1977), form and meaning are subject to their own laws of historical change or resistance to change.

Within realizational theories, two main approaches have been developed, *Realizational Morphology* and *Word and Paradigm Morphology*. Realizational Morphology formalizes how bundles of semantic (typically inflectional) features are realized in phonological form by making use of units for stems, stem variants, and the morphs (now named *exponents*) that realize (or express) sets of inflectional or derivational features (see e.g., Stump, 2001). Realizational Morphology is to some extent compatible with localist models in psychology, in that the stems and exponents of realizational morphology can be seen as corresponding to the "morphemes" (now understood strictly as form-only units, henceforth "morphs") in localist networks. The compatibility is only partial, however, as current localist models typically remain underspecified as to how, in comprehension, the pertinent semantic feature bundles are activated once the proper exponents have been identified. For instance, in the localist interactive activation model of Veríssimo (2018)[4], the exponent `-er` that is activated by the form *teacher* has a connection to a lemma node for ER as deverbal nominalization, but no link is given from the `-er` exponent to an inflectional function that in English is also realized with `-er`, namely, the comparative. Furthermore, even the node for deverbal -ER is semantically underspecified, as -ER realizes a range of semantic functions, including AGENT, INSTRUMENT, CAUSER, and PATIENT (Booij, 1986; Bauer et al., 2015).

A further, empirical, problem for decompositional theories that take the first step in lexical processing to be driven by units for morphs are experiments indicating that quantitative measures tied to properties of whole words, rather than their component morphs, are predictive much earlier in time than expected. For eye-tracking studies on Dutch and Finnish, see Kuperman et al. (2008, 2009, 2010) and for reaction times analyzed with survival analysis, see Schmidtke et al. (2017). These authors consistently find that measures linked to whole words are predictive for shorter response times, and that measures linked to morphs are predictive for longer response times. This strongly suggests that properties of whole words determine early processing and properties of morphs arise later in processing.

There is a more general problem specifically with models that make use of localist networks and the mechanism of interactive activation to implement lexical access. First of all, interactive activation is a very expensive mechanism, as inhibitory connections between morpheme nodes grow quadratically with the number of nodes, and access times increase polynomially or even exponentially. Furthermore, interactive activation as a method for candidate selection in a what amounts to a straightforward classification task is unattractive as it would have to be implemented separately for each classification task that the brain has to carry out. Redgrave et al. (1999) and Gurney et al. (2001) therefore propose a central single mechanism, supposed to be carried out by the basal ganglia, that receives a probability distribution of alternatives as input from any system requiring response selection, and returns the best-supported candidate (see Stewart et al., 2012, for an implementation of their algorithm with spiking neurons).

The second main approach within morpheme-free theories, *Word and Paradigm Morphology*, rejects the psychological reality of stems and exponents, and calls upon proportional analogies between words to explain how words are produced and comprehended (Matthews, 1974; Blevins, 2003, 2006, 2016). Although attractive at a high level of abstraction, without computational implementation, supposed proportional analogies within paradigms do not generate quantitative predictions that can be tested experimentally. As discussed in detail by Baayen et al. (2019), discrimination learning provides a computational formalization of Word and Paradigm Morphology that does generate testable and falsifiable predictions.

In what follows, we will use NDL to estimate a distribution of activations (which, if so desired, could be transformed into probabilities using softmax) over the set of possible word meanings given the visual input. Specifically, we investigate whether we can predict how prior presentation of a prime word affects the activation of the target meaning.

3.2. Morphological Processing Without Morphemes

Naive discriminative learning is not the first cognitive computational model that seeks to move away from morphemes. The explanatory adequacy of the morpheme for understanding lexical processing has also been questioned within psychology by the parallel distributed processing programme (McClelland and Rumelhart, 1986; Rumelhart and McClelland, 1986). As mentioned previously, the triangle model (Harm and Seidenberg, 2004) has been argued to explain the effects of semantic transparency observed for English derived words as reflecting the convergence of phonological and semantic codes (Plaut and Gonnerman, 2000; Gonnerman and Anderson, 2001; Gonnerman et al., 2007). It is noteworthy, however, that to our knowledge, actual simulation studies demonstrating this have not been forthcoming. Importantly, if indeed the triangle model

[4]To our knowledge, this poster presentation is the only computational study addressing morphological processing that makes use of the interactive activation framework.

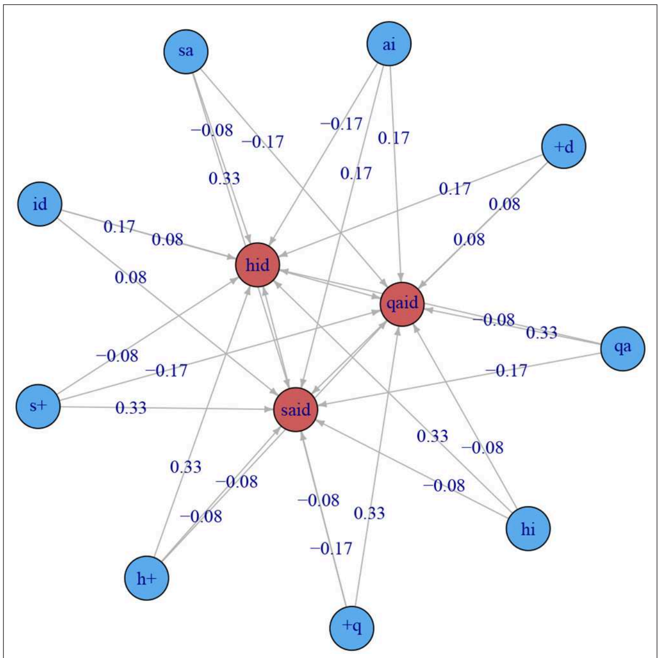


FIGURE 3 | The two-layer network corresponding to $\boldsymbol{W}$. The outer vertices in blue represent the input nodes, the inner vertices in red represent the output nodes. All input nodes are connected to each of the three output nodes. Each edge in the graph is associated with a weight which specifies the support that an input node provides for an output node.

makes correct predictions for English, then one would expect its predictions for German to be wrong, because it would predict semantic transparency effects and no priming for semantically opaque word pairs.

Like the PDP programme, the twin theories of Naive Discriminative Learning (NDL Baayen et al., 2011, 2016b; Sering et al., 2018b) and Linear Discriminative Learning (Baayen et al., 2018, 2019), eschew the construct of the morpheme. But instead of using backpropagation multi-layer networks, NDL and LDL build on simple networks with input units that are fully connected to all output units.

An NDL network is defined by its weight matrix $\boldsymbol{W}$. By way of example, consider the following weight matrix,

$$\boldsymbol{W} = \begin{array}{c} \\ \texttt{\#q} \\ \texttt{qa} \\ \texttt{ai} \\ \texttt{id} \\ \texttt{d\#} \\ \texttt{\#s} \\ \texttt{sa} \\ \texttt{\#h} \\ \texttt{hi} \end{array} \begin{array}{c} \begin{array}{ccc} \textsc{qaid} & \textsc{said} & \textsc{hid} \end{array} \\ \left(\begin{array}{rrr} 0.33 & -0.17 & -0.08 \\ 0.33 & -0.17 & -0.08 \\ 0.17 & 0.17 & -0.17 \\ 0.08 & 0.08 & 0.17 \\ 0.08 & 0.08 & 0.17 \\ -0.17 & 0.33 & -0.08 \\ -0.17 & 0.33 & -0.08 \\ -0.08 & -0.08 & 0.33 \\ -0.08 & -0.08 & 0.33 \end{array}\right) \end{array},$$

which is visualized in **Figure 3**. The output nodes are on the inner circle in red, and the input nodes in the outer circle in blue. A star layout was chosen in order to guarantee readability of the connection weights. The network corresponding to this weight matrix comprises nine sublexical input units, shown in the left margin of the matrix. We refer to these units, here the letter bigrams of the words *qaid*, *said*, and *hid*, as cues; the # symbol (a + in **Figure 3** represents the space character). There are three output units, the outcomes, shown in the upper margin of the matrix. The entries in the matrix present the connection strengths of the digraphs to the lexical outcomes. The digraph `qa` provides strong support (0.33) for QAID ("tribal chieftain"), and `sa` provides strong support (0.33) for SAID. Conversely, `ai`, which is a valid cue for two words, QAID and SAID, has connection strengths to these lexomes of only 0.17. The weights from `hi` and `sa` to QAID are negative, −0.08 and −0.17, respectively. For QAID, the cue that best discriminates this word from the other two words is `qa`. Conversely, `sa` is a (somewhat less strong) discriminative cue arguing against QAID. Informally, we can say that the model concludes the outcome must be QAID given `qa`, and that the outcome cannot be QAID given `sa`.

In the present example, form cues are letter pairs, but other features have been found to be effective as well. Depending on the language and its writing conventions, larger letter or phone n-grams (typically with $1 < n \leq 4$) may outperform letter bigrams. For auditory comprehension, low-level acoustic features have been developed for modeling auditory comprehension (Arnold et al., 2017; Shafaei Bajestan and Baayen, 2018; Baayen et al., 2019). For visual word recognition, low-level visual "histograms of gradient orientation" features have been applied successfully in (Linke et al., 2017).

The total support a_j for an outcome j given the set of cues $\mathcal{C}$ in the visual input to the model, henceforth its activation, is obtained by summing the weights on the connections from these cues to that outcome:

$$a_j = \sum_{i \in \mathcal{C}} w_{ij}.$$

For QAID ($j = 1$), the total evidence a_1 given the cues `#q, qa, ai, id`, and `d#` is $0.33 + 0.33 + 0.17 + 0.08 + 0.08 = 1$.
The values of the weights are straightforward to estimate. We represent the digraph cues of the words by a matrix $\boldsymbol{C}$, with a 1 representing the presence of a cue in the word, and a 0 its absence:

$$\boldsymbol{C} = \begin{array}{c} \\ \textsc{qaid} \\ \textsc{said} \\ \textsc{hid} \end{array} \begin{array}{c} \begin{array}{ccccccccc} \texttt{\#q} & \texttt{qa} & \texttt{ai} & \texttt{id} & \texttt{d\#} & \texttt{\#s} & \texttt{sa} & \texttt{\#h} & \texttt{hi} \end{array} \\ \left(\begin{array}{ccccccccc} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{array}\right) \end{array}.$$

We also represent the outcomes using a matrix, again using binary coding:

$$\boldsymbol{T} = \begin{array}{c} \\ \textsc{qaid} \\ \textsc{said} \\ \textsc{hid} \end{array} \begin{array}{c} \begin{array}{ccc} \textsc{qaid} & \textsc{said} & \textsc{hid} \end{array} \\ \left(\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right) \end{array}.$$

The vectors representing the outcomes in an NDL network are orthogonal: each pair of row vectors of $\boldsymbol{T}$ is uncorrelated. The weight matrix $\boldsymbol{W}$ follows by solving[5]

$$\boldsymbol{CW} = \boldsymbol{T}.$$

In other words, $\boldsymbol{W}$ projects words' forms, represented by vectors in a form space $\{\boldsymbol{C}\}$, onto words' meanings, represented by vectors in a semantic space $\{\boldsymbol{T}\}$[6].

The outcomes of an NDL network represent lexical meanings that are discriminated in a language. Milin et al. (2017a) refer to these outcomes as lexomes, which they interpret as pointers to (or identifiers of) locations (or vectors) in some high-dimensional space as familiar from distributional semantics (see e.g., Landauer and Dumais, 1997; Mikolov et al., 2013). However, as illustrated above with the $\boldsymbol{T}$ matrix, NDL's lexomes can themselves be represented as high-dimensional vectors, the length k of which is equal to the total number of lexomes. The vector for a given lexome has one bit on and all other bits off (cf. Sering et al., 2018b). Thus, the lexomes jointly define a k-dimensional orthonormal space.

However, the orthonormality of the outcome space does not do justice to the fact that some lexomes are more similar to each other than others. Within the general framework of NDL, such similarities can be taken into account, but to do so, measures gauging semantic similarity have to be calculated from a separate semantic space that constructs lexomes' semantic vectors (known as word embeddings in computational linguistics) from a corpus. A technical complication is that, because many words share semantic similarities, the dimensionality of NDL's semantic space, k, is much higher than it need be. As a consequence, the classification accuracy of the model is lower than it could be (see Baayen et al., 2019, for detailed discussion).

The twin model of NDL, LDL, therefore replaces the one-hot encoded semantic vectors as exemplified by $\boldsymbol{T}$ by real-valued vectors. For the present example, this amounts to replacing $\boldsymbol{T}$ by a matrix, such as $\boldsymbol{S}$:

$$\boldsymbol{S} = \begin{array}{c} \\ \text{QAID} \\ \text{SAID} \\ \text{HID} \end{array} \begin{array}{c} \begin{array}{ccc} \text{QAID} & \text{SAID} & \text{HID} \end{array} \\ \begin{pmatrix} 0.4 & -0.2 & 0.3 \\ -0.2 & -0.2 & -0.3 \\ -0.1 & 0.3 & 0.3 \end{pmatrix} \end{array}.$$

Actual corpus-based semantic vectors are much longer than this simple example suggests, with hundreds or even thousands of elements. The method implemented in Baayen et al. (2019) produces vectors the values of which represent a given lexome's collocational strengths with all the other lexomes in the corpus.

Model accuracy is evaluated by examining how close a predicted semantic vector $\hat{\boldsymbol{s}}$ is to the targeted semantic vector $\boldsymbol{s}$, a row vector of $\boldsymbol{S}$. In the case of NDL, this evaluation is straightforward: The lexome that is best supported by the form features in the input, and that thus receives the highest activation, is selected. In the case of LDL, that word ω is selected as recognized for which the predicted semantic vector $\hat{\boldsymbol{s}}$ is most strongly correlated with the targeted semantic vector $\boldsymbol{s}_{\omega}$.

As the dimensionality of the row vectors of $\boldsymbol{S}$, $\boldsymbol{T}$, and $\boldsymbol{C}$ can be large, with thousands or tens of thousands of values, we refer to the network $\boldsymbol{W}$ as a "wide learning" network, as opposed to "deep learning" networks which have multiple layers but usually much smaller numbers of units on these layers.

Of specific relevance to the present study is how NDL and LDL deal with morphologically complex words. With respect to the forms of complex words, exactly the same encoding scheme is used as for simple words, with either n-grams or low-level modality-specific features used as descriptor sets. No attempt is made to find morpheme boundaries, stems, affixes, or allomorphs.

At the semantic level, both NDL and LDL are analytical. NDL couples inflected words, such as *walked* and *swam* with the lexomes WALK and PAST, and SWIM and PAST, respectively. In the example worked out in **Table 3**, the word form *walk* has LX6 as identifier; the lexome for *past* is indexed by LX4. The form *walked* is linked with both LX6 and LX4. For clarity of exposition, instead of using indices, we refer to lexomes using small caps: WALK and PAST. An NDL network is thus trained to predict, for morphologically complex words, on the basis of the form features in the input, the simultaneous presence of two (or more) lexomes. Mathematically, as illustrated in the top half of **Table 3**, this amounts to predicting the sum of the one-hot encoded vectors for the stem (WALK) and the inflectional function (PAST). Thus, NDL treats the recognition of complex words as a multi-label classification problem (Sering et al., 2018b).

LDL proceeds in exactly the same way, as illustrated in the bottom half of **Table 3**. Again, the semantic vector of the content lexome and the semantic vector of the inflection are added. The columns now label semantic dimensions. In the model of Baayen et al. (2019), these dimensions quantify collocational strengths with—in the present example—10 well-discriminated lexomes. Regular past tense forms, such as *walked* and irregular past tense forms, such as *swam* are treated identically at the semantic level. It is left to the mapping $\boldsymbol{W}$ (the network taking form vectors as input and producing semantic vectors as output) to ensure that the different forms of regular and irregular verbs are properly mapped on the pertinent semantic vectors.

The NDL model as laid out by Baayen et al. (2011) treats transparent derived words in the same way as inflections, but assigns opaque derived words their own lexomes. For opaque words in which the semantics of the affix are present, even though there is no clear contribution from the semantics of the base word, a lexome for the affix is also included (e.g., *employer*: EMPLOY + ER; *cryptic*: CRYPTIC + IC).

The LDL model, by contrast, takes the idea seriously that derivation serves word formation, in the onomasiological sense. Notably, derived words are almost always characterized by semantic idiosyncracies, the exception being inflection-like derivation, such as adverbial *-ly* in English[7]. For instance, the

[5]In R: `W = ginv(C) %*% T`, see Baayen et al. (2018) and Baayen et al. (2019) for further details on linear transformations from form to meaning (and from meaning to form).

[6]The weight matrix $\boldsymbol{W}$ is identical to the weight matrix obtained by applying the equilibrium equations of Danks (2003) for the Rescorla-Wagner learning rule (Rescorla and Wagner, 1972) that was used by Baayen et al. (2011), see Sering et al. (2018b) for detailed discussion. $\boldsymbol{W}$ is also identical to the matrix of beta weights of a multivariate multiple regression model regressing the semantic vectors on the form vectors.

[7]The reason that adverbial *-ly* is generally treated as derivational is that the word category of words with *-ly* is not identical to the word category of its base word.

TABLE 3 | Semantic vector representations for inflected words in NDL (top) and LDL (bottom).

	NDL									
	Lx1	**Lx2**	**Lx3**	**Lx4**	**Lx5**	**Lx6**	**Lx7**	**Lx8**	**Lx9**	**Lx10**
Walk	0	0	0	0	0	1	0	0	0	0
Past	0	0	0	1	0	0	0	0	0	0
Walked	0	0	0	1	0	1	0	0	0	0

	LDL									
	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**
Walk	−0.16	0.25	−0.07	−0.04	0.13	0.03	−0.04	0.10	0.22	−0.25
Past	0.16	0.13	−0.36	0.01	−0.07	−0.04	−0.32	0.07	0.45	0.14
Walked	0.00	0.38	−0.43	−0.03	0.06	−0.01	−0.36	0.17	0.67	−0.11

English word *worker* denotes not just "someone who works," but "one that works especially at manual or industrial labor or with a particular material," a "factory worker," "a member of the working class," or "any of the sexually underdeveloped and usually sterile members of a colony of social ants, bees, wasps, or termites that perform most of the labor and protective duties of the colony" (https://www.merriam-webster.com/dictionary/worker, s.v.). Given these semantic idiosyncracies, when constructing semantic vectors from a corpus, LDL assigns each derived word its own lexome. However, in order to allow the model to assign an approximate interpretation to unseen derived words, each occurrence of a derived word is also coupled with a lexome for the semantic function of the affix. For instance, *worker* (in the sense of the bee) is associated with the lexomes WORKER and AGENT, and *amplifier* with the lexomes AMPLIFIER and INSTRUMENT. In this way, semantic vectors are created for derivational functions, along with semantic vectors for the derived words themselves (see Baayen et al., 2019, for detailed discussion and computational and empirical evaluation)[8].

To put LDL and NDL in perspective, consider the substantial advances made in recent years in machine learning and its applications in natural language engineering. Computational linguistics initially worked with deterministic systems applying symbolic units and formal grammars defined over these units. It then became apparent that considerable improvement in performance could be obtained by making these systems probabilistic. The revolution in machine learning that has unfolded over the last decade has made clear that yet another substantial step forward can be made by moving away from hand-crafted systems building on rules and representations, and to make use instead of deep learning networks, such as autoencoders, LSTM networks for sequence to sequence modeling, and deep convolutional networks, outperforming almost all classical symbolic algorithms on tasks as diverse as playing *Go* (AlphaGo, Silver et al., 2016) and speech recognition (deep speech, Hannun et al., 2014). How far current natural language processing technology has moved away from concepts in classical (psycho)linguistics theory is exemplified by Hannun et al. (2014), announcing in their abstract that they "…do not need a phoneme dictionary, nor even the concept of a 'phoneme"' (p. 1).

The downside of the algorithmic revolution in machine learning is that what exactly the new networks are doing often remains a black box. What is clear, however, is that these networks are sensitive to what in regression models would be higher-order non-linear interactions between predictors (Cheng et al., 2018). Crucially, such complex interactions are impossible to reason through analytically. As a consequence, models for lexical processing that are constructed analytically by hand-crafting lexical representations for stems and exponents, and hand-crafting inhibitory or excitatory connections between these representations, as in standard interactive activation models, are unable to generate sufficiently accurate estimates for predicting with precision aspects of human lexical processing.

We note here that NDL and LDL provide high-level functional formalizations of lexical processing. They should not be taken as models for actual neural processing: biological neural networks involve cells that fire stochastically, with connections that are stochastic (Kappel et al., 2015, 2017) as well. Furthermore, most neural computations involve ensembles of spiking neurons (Eliasmith et al., 2012).

NDL and LDL are developed to provide a linguistically fully interpretable model using mathematically well-understood networks that, even though very simple, are powerful enough to capture important aspects of the interactional complexities in language, and to generate predictions that are sufficiently precise to be pitted against experimental data. Although NDL and LDL make use of the simplest possible networks, these networks can, in combination with carefully chosen input features, be surprisingly effective. For instance, for auditory word recognition, an NDL model trained on the audio of individual words extracted from 20 hours of German free conversation correctly recognized around 20% of the words, an accuracy that was subsequently found to be within the range of human recognition accuracy (Arnold et al., 2017). Furthermore, Shafaei Bajestan and Baayen (2018) observed that NDL outperforms deep speech networks by a factor 2 on isolated word recognition. With respect to visual word recognition, Linke et al. (2017) showed, using low-level visual features, that NDL outperforms deep convolutional networks (Hannagan et al., 2014) on the task of predicting word learning in baboons (Grainger et al., 2012). For a systematic comparison of NDL/LDL with interactive activation and parallel distributed processing approaches, the reader is referred to **Appendix B**.

3.3. Computational Modeling With Naive Discriminative Learning

In the present study, we model our experiment with NDL, rather than LDL, for two reasons. First, it turns out that NDL, the simpler model, is adequate. Second, work is in progress to derive corpus-based semantic vectors for German along the lines of Baayen et al. (2019), which will include semantic vectors for inflectional

[8]For a compositional approach to the semantics of complex words using distributional semantics, see Lazaridou et al. (2013) and Marelli and Baroni (2015).

and derivational semantic functions, but these vectors are not yet available to us.

The steps in modeling with NDL are the following. First, the data on which the network is to be trained have to be prepared. Next, the weights on the connections from the form features to the lexomes are estimated. Once the network has been trained, it can be used to generate predictions for the magnitude of the priming effect. In the present study, we generate these predictions by inspecting the extent to which the form features of the prime support the lexome of the target.

3.3.1. Data Preparation

The data on which we trained our NDL network comprised 18,411 lemmas taken from the CELEX database, under the restrictions that (i) they contained no more than two morphemes according to the CELEX parses, (ii) that the word was not a compound, and (iii) that it either had a non-zero CELEX frequency or occurred as a stimulus in the experiment. One stimulus word, *betraten*, was not listed in CELEX, and hence this form was not included in the simulation study. For each lemma, its phonological representation and its frequency were retrieved from CELEX. As form cues, we used triphones (for the importance of the phonological route in reading, see Baayen et al., 2019, and references cited there).

Each lemma was assigned its own lexome (but homophones were collapsed). The decision to assign each lemma its own lexome follows Baayen et al. (2019) and departs from Baayen et al. (2011). This similar treatment of transparent and opaque verbs is motivated by several theoretical considerations. First, there is no binary distinction between transparent and opaque. The meanings of particle verbs lie on a continuum between relatively semantically compositional and relatively semantically opaque. Second, even the compositional interpretation of a supposedly transparent verb, such as *aufstehen* ("stand up") is not straightforward in the absence of situational experience—the particle *auf* (roughly meaning *on* or *onto*) may express a wide range of meanings, depending on cotext and context. In what follows, we therefore assume that even transparent complex words possess somewhat idiosyncratic meanings, and hence should receive their own lexomes in the NDL network.

The resulting input to the model was a file with 4,492,525 rows and two columns, one column spelling out a word's triphones, and the other column listing its lexome. Each word appeared in the file with a number of tokens equal to its frequency in CELEX. The order of the words in the file was randomized.

3.3.2. Training the Network

An NDL network with 10,180 input nodes (triphones) and 18,404 output nodes (lexomes) was trained on the input list, with incremental updating of the weights on the connections from features in the input to the lexomes, using the learning rule of Rescorla and Wagner (1972) ($\lambda = 1, \alpha = 0.001, \beta = 1$; i.e., with a learning rate of 0.001). As there were 4,492,525 learning events in the input file, the total number of times that weights were updated was 4,492,525[9].

3.4. Modeling Priming

To model the effect of priming, we presented the triphones of the prime to the network, and summed the weights on the connections from these triphones to the pertinent target to obtain a measure of the extent to which the prime pre-activates its target (henceforth `PrimeToTargetPreActivation`). **Figure 4**, upper left panel, presents a boxplot for `PrimeToTargetPreActivation` as a function of `PrimeType`. Interestingly, the opaque and transparent prime types comprise prime-target pairs for which the prime provides substantial and roughly the same amount of pre-activation for the target. For the other prime types, pre-activation is close to zero. Form-related prime-target pairs show some pre-activation, but this pre-activation is much reduced compared to the prime-target pairs in the opaque and transparent conditions.

The upper right panel of **Figure 4** presents the results obtained when the empirical frequencies with which words were presented to the NDL network are replaced by uniformly distributed frequencies. This type-based simulation generates predictions that are very similar to those of the token-based simulation. This result shows that imprecisions in the frequency counts underlying the token-based analysis are not responsible for the model's predictions.

Above, we called attention to the finding of Smolka and Eulitz (2011) that very similar priming effects are seen when the order of prime and target is reversed. We therefore also ran a simulation in which we reversed the order of prime and target, and investigated the extent to which the current targets (now primes) co-activate the current primes (now targets). The distributions of the predicted pre-activations are presented in the lower left panel of **Figure 4** (`target-to-prime pre-activation`). Apart from one extreme outlier for the opaque condition, the pattern of results is qualitatively the same as for the Prime-to-Target Pre-Activation presented in the upper panel. For both simulations, there is no significant difference in the mean for the opaque and transparent conditions, whereas these two conditions have means that are significantly larger than those of the other three condition (Wilcoxon-tests with Bonferroni correction). In summary, our NDL model generates the correct prediction that the priming effect does not depend on the order of prime and target.

Reaction times are expected to be inversely proportional to `PrimeToTargetPreActivation`. We therefore ran a linear model on the stimuli, and used the reciprocal of `PrimeToTargetPreActivation` as response variable, based on the simulation in which the model was presented with the empirical word frequencies. As the resulting distribution is highly skewed, the response variable was transformed to $\log(1/(\texttt{PrimeToTargetPreActivation} + 0.14)$[10]. The opaque and transparent priming conditions were supported as having significantly shorter simulated reaction times compared

[9]Optimized software, e.g., Sering et al. (2017), makes it possible to harness multiple cores in parallel. Using 6 cores, training the network takes <10 min. Incremental learning is much faster than weight estimation by means of the Danks equilibrium equations, which were used by Baayen et al. (2011).

[10]The shift 0.14 is slightly larger than the absolute value of the most negative pre-activation. This shift thus ensures that all pre-activation values are positive, so that a log-transform becomes possible.

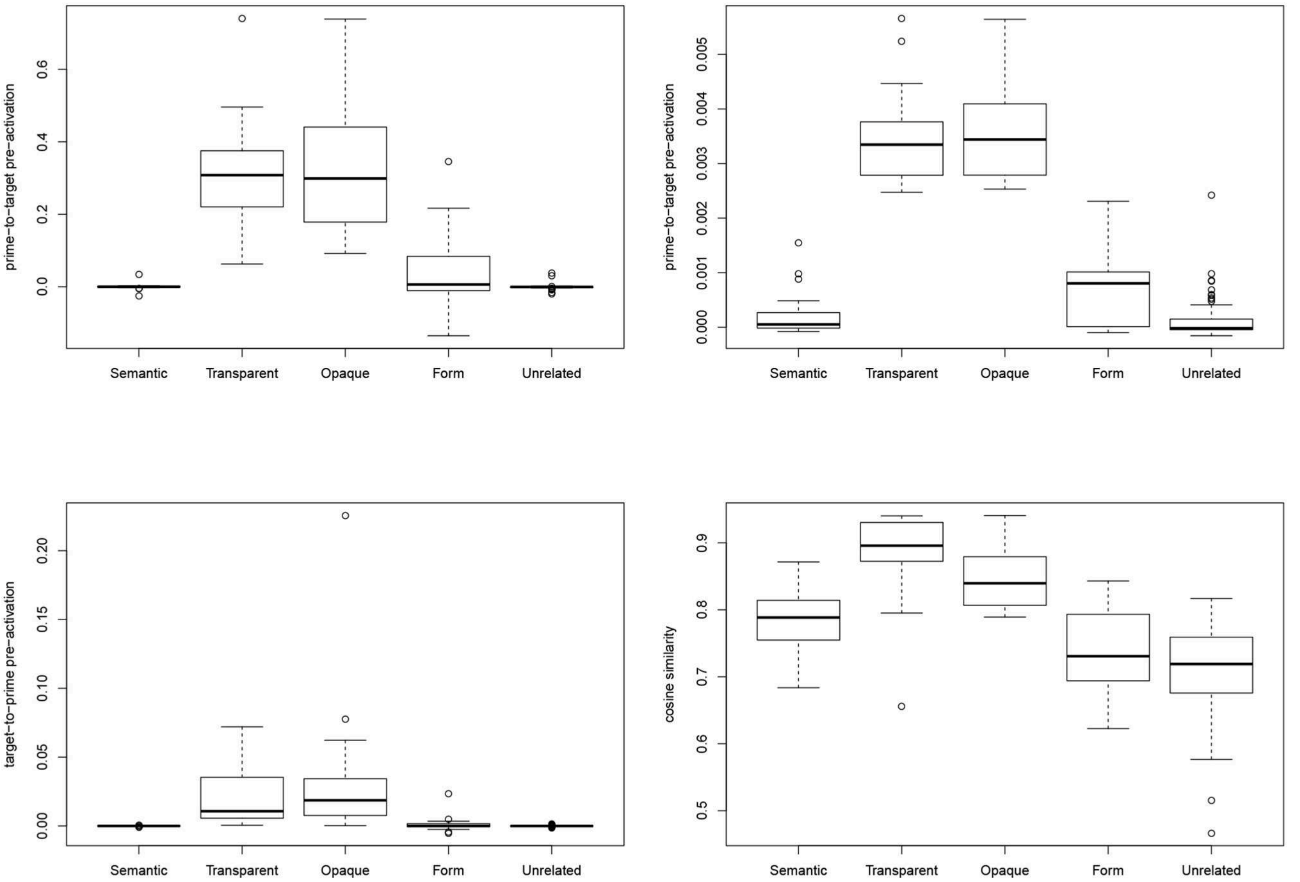


FIGURE 4 | Predicted NDL Prime-to-Target Pre-Activation using empirical word frequencies (upper left panel), Predicted NDL Prime-to-Target Pre-Activation using uniformly distributed frequencies (upper right panel), Target-to-Prime Pre-Activation using empirical frequencies (lower left panel), and Prime-target Cosine Similarity (lower right panel), broken down by prime type.

to the unrelated condition (both $p \ll 0.0001$), in contrast to the other two conditions (both $p > 0.5$).

Recall that the outcome vectors of NDL are orthogonal, and that hence the present NDL models all make predictions that are driven purely by form similarity. The model is blind to potential semantic similarities between primes and targets, not only for the primes and targets in the transparent and opaque conditions, but also to semantic similarities present for the other prime types. To understand to what extent semantic similarities might be at issue in addition to form similarities, we therefore inspected prime and target's semantic similarity using distributional semantics.

3.5. Semantic Vectors From Tweets

As LDL-based semantic vectors for German are currently under construction, we fell back on the word embeddings (semantic vectors) provided at http://www.spinningbytes.com/resources/wordembeddings/ (Cieliebak et al., 2017; Deriu et al., 2017). These embeddings (obtained with `word2vec`, Mikolov et al., 2013) are 300-dimensional vectors derived from a 50 million word corpus of German tweets. Tweets are relatively short text messages that reflect spontaneous and rather emotional conversation. Tweets from facebook have been shown to outperform frequencies from standard text corpora in predicting lexical decision latencies (Herdağdelen and Marelli, 2017).

Cieliebak et al. (2017) and Deriu et al. (2017) provide separate semantic vectors for words' inflected variants. For instance, the particle verb *vorwerfen* ("accuse") occurs in their database in the forms *vorwerfen* (infinitive and 1st or 3rd person plural present), *vorwerfe* (1st person singular present), *vorwirfst* (2nd person singular present), *vorwirft* (3rd person singular present), *vorwerft* (2nd person plural present), *vorgeworfen* (past participle), and *vorzuwerfen* (infinitive construction with *zu*). As we can expect for tweets, not all inflected forms, in particular the more formal ones, appear in the database. Importantly, the semantic vectors are probably obtained without taking into account that the particle of a particle verb can appear separated from its verb, sometimes at a considerable distance (see Schreuder, 1990, for discussion of the cognitive consequences of this separation), as in the sentence "Sie *wirft* ihm seinen Leichtsinn *vor*," "She accuses him of his thoughtlessness." Given the computational complexity of identifying particle-verb combinations when the particle appears at a distance, it is highly likely that for split particle verbs, the base verb of the verb-particle combination is processed as if it were a simple verb (e.g., *werfe*, *wirfst*, *wirft*,

TABLE 4 | Effect of PrimeType in a linear model predicting cosine similarity, using treatment coding with the semantic condition as reference level.

	Estimate	Std. Error	*t*-value	Pr (>|t|)
Intercept (Semantic)	0.7778	0.0134	58.19	0.0000
PrimeType = Form	−0.0404	0.0197	−2.05	0.0418
PrimeType = Opaque	0.0696	0.0191	3.64	0.0004
PrimeType = Transparent	0.1083	0.0191	5.66	0.0000
PrimeType = Unrelated	−0.0700	0.0150	−4.67	0.0000

TABLE 5 | GAMM fitted to inverse transformed reaction times using model-based predictors; te(): tensor product smooth.

A. Parametric coefficients	Estimate	Std. Error	*t*-value	*p*-value
Intercept	−1.9404	0.0294	−66.0214	< 0.0001
B. Smooth terms	**edf**	**Ref.df**	***F*-value**	***p*-value**
te(Target Activation × Prime-to-target pre-activation × Prime-to-target cosine similarity)	14.9326	17.4660	10.3959	<0.0001
Random intercepts prime	15.4237	108.0000	0.1888	0.0547
Factor smooths for trial × subject	101.6786	449.0000	2.4620	<0.0001
Random intercepts target	19.9620	40.0000	1.2681	<0.0001

werfen, and *werft*, 1st, 2nd, and 3rd person singular and plural present, respectively). As a consequence, the semantic similarity of simple verbs and particle verbs computed from the word embeddings provided by Cieliebak et al. (2017) and Deriu et al. (2017) is in all likelihood larger than it should be.

Not all words in the experiment are in this database; but for six words, we were able to replace the infinitive by a related form (*einpassen* → *reinpassen*, *verqualmen*→ *verqualmt*, *fortlaufen* → *fortlaufend*, *bestürzen* → *bestürzend*, *verfinstern* → *verfinstert*, *beschneien* → *beschneites*).

For each prime-target pair for which we had data, we calculated the cosine similarity of the semantic vectors of prime and target, henceforth `Prime-to-Target Cosine Similarity`. **Figure 4**, lower right panel, shows that the transparent pairs have the greatest semantic similarity, followed by the opaque pairs, then the semantic pairs and the form pairs, and the least semantic similarity by the unrelated pairs.

Surprisingly, the semantic controls have a rather low semantic similarity, substantially less than that of the opaque pairs. A linear model with the semantic primes as reference level clarifies that the semantic pairs are on a par with the form controls, more similar than the unrelated pairs, but less similar than both the opaque and transparent pairs (**Table 4**).

There is a striking discrepancy between the assessment of semantic similarity across prime types based on the cosine similarity of the semantic vectors on the one hand, and an assessment based on the ratings for semantic relatedness between word pairs, as documented in **Table 1**. In the former, semantic pairs pattern with form controls and differ from transparent ones, while in the latter, semantic pairs pattern with transparent pairs (5.5 and 5.7, respectively), and opaque pairs with form-related ones (2.1 and 1.7, respectively).

Most important to our study is that the opaque pairs show significantly less semantic similarity than the transparent ones ($p < 0.0047$, Wilcoxon test): The analysis of word embeddings confirms that there is a true difference in semantic transparency between the transparent and opaque prime-target pairs. And yet, this difference is not reflected in our reaction times.

Given the strong track record of semantic vectors in both psychology and computational linguistics, the question arises of whether the prime-target cosine similarities are predictive for reaction times, and how the magnitude of their predictivity compares to that of the NDL Prime-to-Target Pre-Activation.

3.6. Putting It All Together: Predicting Reaction Times

To address these questions, we fitted a new GAMM to the inverse-transformed reaction times of our experiment, replacing the factorial predictor `PrimeType` with the model-based predictor `Prime-to-Target Pre-Activation`. We also replaced target frequency by the activation that the target word receives from its own triphones (henceforth `TargetActivation`, see Baayen et al. (2011) for detailed analyses using this measure). Target activation is proportional to frequency, and hence larger values of target activation are expected to indicate shorter response times[11].

Of the experimental dataset, about 7% of the observations was lost due to 7 words not being available in CELEX or in the dataset of word embeddings. To set a baseline for model comparison, we refitted the GAMM discussed above to the 1999 datapoints of the reduced dataset. The fREML score for this model was 360.76. A main effects model replacing `Target Frequency` by `Target Activation`, `PrimeType` by `Prime-to-Target Pre-Activation`, and as additional predictor the `Prime-to-Target Cosine Similarity` had a slightly higher fREML score, 370.97. An improved model was obtained by allowing the three new covariates to interact, using a tensor product smooth. The fREML score of this model, summarized in **Table 5**, was 354.16. A chi-squared test for model comparison (implemented in the `compareML` function of the **itsadug** package van Rij et al., 2017) suggests that the investment of 4 additional effective degrees of freedom is significant ($p = 0.010$). As the models are not nested and the increase in goodness of fit is moderate (6.6 fREML units), we conclude that—to obtain a model that is at least equally good—it is possible to replace the classical predictors, such as Frequency and PrimeType by model-based predictors without loss of prediction accuracy[12].

[11] Both activation measures were log-transformed after adding a small number, 0.14 for `Prime-to-Target Pre-Activation` and 0.01 for `Target Activation`, to ensure that all pertinent numbers were positive before taking logarithms.

[12] Including an interaction of PrimeType by Target Frequency in the GAMM with classical predictors led to an increase in the fREML score, indicating overfitting and increased model complexity without increased prediction accuracy.

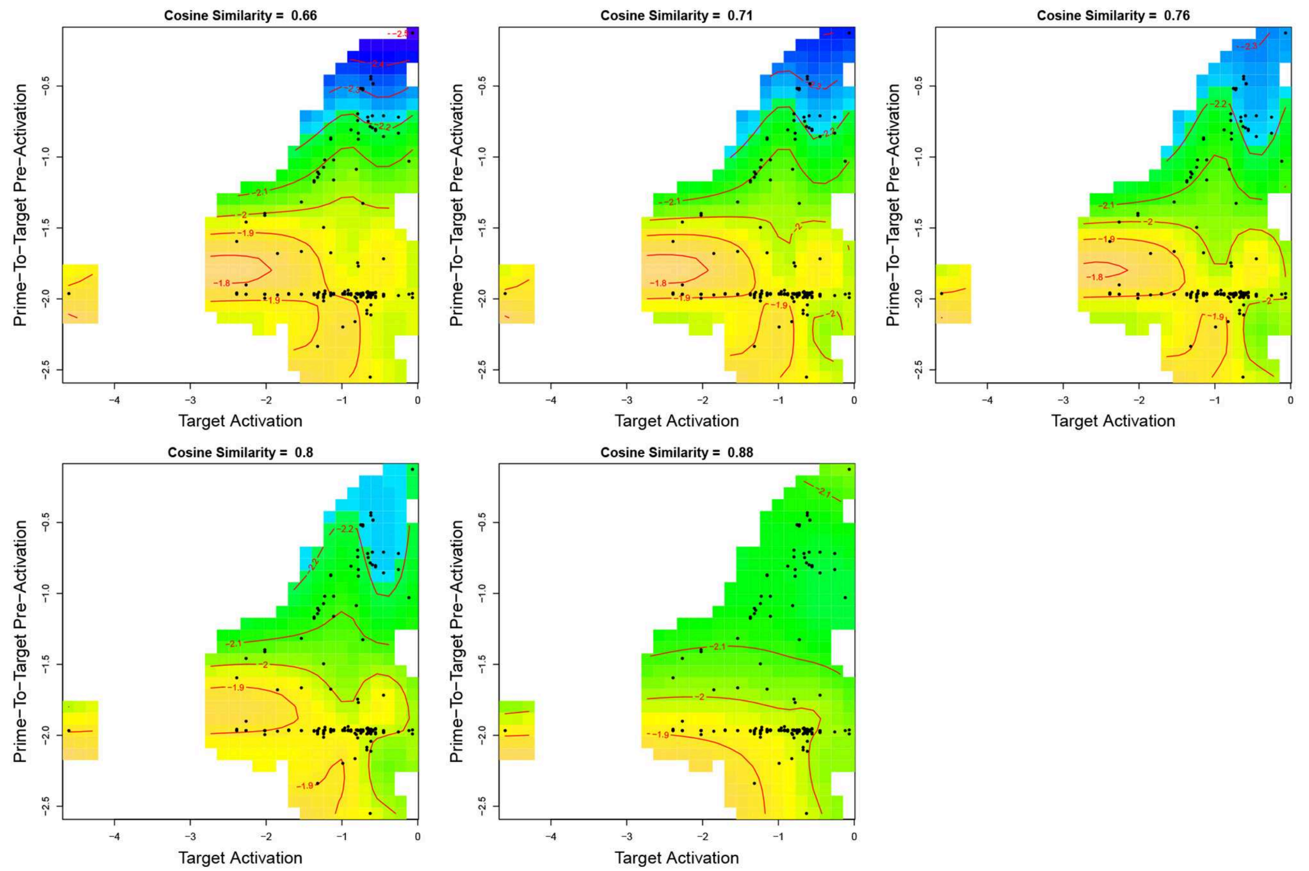


FIGURE 5 | Interaction of Target Activation by Prime-to-Target Pre-Activation by Prime-to-Target Cosine Similarity in a GAMM fitted to the inverse-transformed reaction times. Darker shades of blue indicate shorter reaction times, and darker shades of yellow longer RTs. For visualization, one extreme outlier value for Prime-to-Target Pre-Activation was removed (13 datapoints). Contour lines are 0.1 inverse RT units apart.

The three-way interaction involving `Target Frequency`, `Prime-To-Target Pre-Activation`, and `Prime-to-Target Cosine Similarity` is visualized in **Figure 5**. The five panels show the joint effect of the first two predictors for selected quantiles of `Prime-to-Target Cosine Similarity`: From top left, to bottom right, Prime-to-Target Cosine Similarity is set to its 0.1, 0.3, 0.5, 0.7, and 0.9 deciles. Darker shades of blue indicate shorter reaction times, and darker shades of yellow longer RTs.

As can be seen in the upper left panel (for the first decile of `Cosine Similarity`), reaction times decrease slightly as `Target Activation` is increased, but only when there is little `Prime-To-Target Pre-Activation`. A clear effect of `Prime-to-Target Pre-Activation` is present for the larger values of `Target Activation`.

Recall that, as shown in **Figure 4**, transparent and opaque prime-target pairs have the same mean pre-activation, whereas the mean cosine similarity is greater for transparent prime-target pairs compared to opaque pairs. If both pre-activation and cosine similarity would have independent effects, one would expect a difference in the mean reaction times for these two prime types, contrary to fact. The interaction of pre-activation by target activation by cosine similarity resolves this issue by decreasing the effect of pre-activation as cosine similarity increases. When prime and target are more similar semantically, the effect of pre-activation is reduced, and reaction times are longer than would otherwise have been the case. This increase in RTs may reflect the cognitive system slowing down to deal with two signals for very similar meanings being presented in quick succession in a way that is extremely rare in natural language.

We checked whether the association ratings that were used for stimulus preparation were predictive for the reaction times. This turned out not to be the case, not for the model with classical predictors, nor when the association ratings were added to the model with discrimination-based predictors.

Finally, **Table 6** presents the fREML scores for the full model, and the three models obtained when one covariate is removed at a time. Since smaller fREML scores indicate a better fit, **Table 6** clarifies that `Prime-to-Target Pre-Activation` is the most important covariate, as its exclusion results in the worst model fit (386.11). The variable importances of `Prime-to-Target Cosine Similarity` is also substantial (383.49) whereas removing `Target Activation` from the model specification reduces

TABLE 6 | fREML scores for four models: the full model, a model without Prime-to-Target Cosine Similarity, a model without Prime-to-Target Pre-Activation, and a model without Target Activation.

Model	fREML score
Full model	354.16
te(Target activation × prime-to-target pre-activation)	383.49
te(Target activation × prime-to-target cosine similarity)	386.11
te(Prime-to-target pre-activation × prime-to-target cosine similarity)	361.62

A smaller fREML score indicates a better fit. te(), tensor product smooth.

model fit only slightly (361.62). The model that best predicts (AIC 354.16) pure morphological priming includes all three factors: `Prime-to-Target Pre-Activation` (capturing Prime Type), `Prime-to-Target Cosine Similarity` (capturing minor semantic effects), and `Target Activation` (capturing the frequency effect).

4. DISCUSSION

We presented an overt primed visual lexical decision experiment that replicated earlier results for German complex verbs: Priming effects were large and equivalent for semantically transparent and semantically opaque prime-target pairs.

These findings add to the cumulative evidence of "pure morphological priming" patterns that suggest stem access independent of semantic compositionality in German, in contrast to English and French, where semantic compositionality has been reported to co-determine word processing.

One could argue that the fact that we used verbs might account for these cross-language differences, in particular because the majority of the verbs in German hold prefixes or particles. Indeed, there are only few studies in English that applied prefixed stimuli that can be interpreted as verbs, such as *preheat–heat* (e.g., Exp. 5 in Marslen-Wilson et al., 1994b; Exp. 4 in Gonnerman et al., 2007; EEG-experiment in Kielar and Joanisse, 2011). In these experiments, though, opaque prime-target pairs like *rehearse–hearse* did not induce priming, which contrasts with our findings in German. Furthermore, in a previous study (Exp. 3 in Smolka et al., 2014), we applied both verbs and nouns or adjectives in the same experiment to make sure that the opaque priming effects are not due to the fact that participants see verbs only. It might indeed be the case that, in German, morphological constituents assume a more prominent role than prefixes and suffixes in nouns and adjectives in languages as English and French, and that the complexity and structure of the language affects its (lexical) processing (Günther et al., 2019).

Most importantly, because neither localist nor connectionist models of lexical processing are able to account for the German findings, Smolka et al. (2009, 2014, 2019) proposed a stem-based frequency account, according to which stems constitute the crucial morphological units regulating lexical access in German, irrespective of semantic transparency.

In the present study, we took the next step and modeled the German stem priming patterns using naive discriminative learning (NDL). This morpheme-free computational model clarifies that the observed priming effects across all prime types may follow straightforwardly from basic principles of discrimination learning. The extent to which sublexical features of the prime (letter triphones) pre-activate the lexome of the target is the strongest predictor for the reaction times. A substantially smaller effect emerged for the activation of the target (comparable to a frequency effect). The semantic similarity of prime and target as gauged by the cosine similarity measure also had a solid effect.

The semantic vectors (word embeddings) used for calculating the cosine similarity between primes and their targets were taken from a database of German tweets. It is noteworthy that the cosine similarity measure provided good support for the transparent prime-target pairs being on average more semantically similar than the opaque prime-target pairs. However, this difference between the two prime types was not reflected in the corresponding mean reaction times.

A three-way interaction between prime-to-target pre-activation, target activation, and prime-target cosine similarity detected by a generalized additive mixed model fitted to the reaction times clarified that as the semantic similarity of primes and targets increases, the facilitatory effect of pre-activation decreases. Apparently, when primes and targets are more similar, pre-activation by the prime forces the cognitive system to slow down in order to resolve the near simultaneous activation of two very similar, but conflicting, meanings.

Interestingly, even though across prime types stimuli were matched for association ratings, these ratings were not predictive for reaction times. Stimuli were not matched across prime types for the cosines of the angles between primes' and targets' tweet-based semantic vectors, yet, surprisingly, these were predictive for reaction times. This finding is particularly surprising for the present data on German, as in previous work semantic similarity measures (not only human but also vector-based measures like LSA (Landauer and Dumais, 1997) and HAL (Lund and Burgess, 1996)) were observed not to be predictive of reaction times. It is conceivable that the present semantic measure is superior to LSA and HAL, due to it being calculated from a large volume of tweets—Herdağdelen and Marelli (2017) point out that measures based on distributional semantics calculated from corpora of social media provide excellent predictivity for lexical processing.

A caveat is in order, though, with respect to the cosine similarity measure, as in all likelihood particle verbs and their simple counterparts are estimated to be somewhat more similar than they should be. Particles can be separated by several words from their stems, and these stems will therefore be treated as simple verbs by the algorithm constructing semantic vectors (especially when the particle falls outside word2vec's 5-word moving window). This, however, implies that the cosine similarity measure must be less sensitive than it could have been: The vector for the simple verb is artifactually shifted in the direction of the vector of its particle verb, with the extent of this shift depending on the frequency of the simple verb, and the frequencies of the separated and non-separated derived particle verbs. As simple verbs typically occur with several particles, the semantic vectors for these simple verbs are likely to have shifted somewhat in the direction of the centroid of the vectors of its particle verbs.

Nevertheless, the present tweet-based semantic vectors contribute to the prediction of reaction times. Importantly, there is no a-priori reason for assuming that the rate at which particles occur separated from their stem would differ across prime types. As a consequence, the partial confounding of particle verbs and simple verbs by the distributional semantics algorithm generating the semantic vectors cannot be the main cause of the different distributions of cosine similarities across the different prime types (in this context, it is worth noting that any skewing in frequency counts of complex and simple verbs does not affect the qualitative pattern of results for the predictor with the greatest effect size, the prime-to-target activation measure, as shown by the simulation in which all words are presented to the network with exactly the same frequency. We leave answering the question of why this particular frequency measure appears to be effective as predictor for the present reaction times to further research).

An important result is that a generalized additive mixed model fitted to the reaction times provides a fit that is at least as good, if not slightly better, when the activation, pre-activation, and cosine similarity measures are used, compared to when prime type and frequency of occurrence are used as predictors (reduction in fREML score: 6.6, which, for 4 degrees of freedom is significant ($p = 0.010$) according to the chi-squared test implemented in the `compareML` function of the **itsadug** package)[13].

The theoretical contribution of this study is that it challenges the general localist interactive activation framework that dominates the current discourse on morphological processing. Stems and morphemes are assumed to be psychologically real (see e.g., Zwitserlood, 2018), and to excite or inhibit each other. Furthermore, these high-level concepts are apparently understood to be sufficient for explaining the effects of experimental manipulations. NDL, by contrast, provides a framework within which quantitative measures can be derived that can be pitted against experimental response variables. NDL (and LDL) make use of the simplest possible network, the mathematics of which are well understood—in essence, NDL is nothing more (or less) than incremental multivariate multiple (logistic) regression. An NDL model is essentially parameter-free[14], and driven completely by the distributional properties of the words in the corpus it is trained on. NDL, just as the interactive activation framework, requires the analyst to make decisions on input and output nodes, but unlike the interactive activation framework, no hand-crafting of connections is required, and no search is required for finding a set of parameters that make the model behave in the way desired. As a consequence, the measures derived from an NDL network can be used simply as a statistical tool for assessing how well a word's meaning can be "classified" or "discriminated" given its form features. As expected, reaction times become shorter when target meanings are better discriminated, i.e., when target activation is higher and the probability of the target being correctly classified is greater. Furthermore, when a morphologically related prime is presented to the network, a distribution of activations over the lexomes ensues in which the activation of the target is greater compared to trials with primes that are not morphologically related. This pre-activation of the target by reading the prime apparently carries over to the reading of the target[15].

Although NDL-based measures can be used in the same way as measures such as word frequency and neighborhood density, the linguistic theory underlying NDL holds that morphemes (in the sense of minimal signs) as well as sublexical form units, such as stems and exponents are not necessary. At the same time, this theory is analytical at the semantic level. What Baayen et al. (2019) and Baayen et al. (2018) have shown for LDL is that accurate mappings between form vectors and distributional semantic vectors can be set up with linear transformations, i.e., with simple two-layer networks (and no hidden layers). The more comprehensive model of the mental lexicon developed in Baayen et al. (2019) and Chuang et al. (2019) makes use of multiple such networks to generate quantitative predictions for auditory comprehension (with audio as input), visual comprehension, and speech production. A proof of concept that inflected forms of rich paradigms can be predicted from their corresponding semantic vectors without requiring sublexical form units, such as stems and exponents is provided by Baayen et al. (2018). It is within this wider context that the present computational modeling of overt primed visual lexical decision comes into its own.

It is important to note that the design decision to assign complex verbs their own lexome, irrespective of semantic transparency, following Baayen et al. (2019), is crucial for enabling NDL to simulate the German behavioral priming data. This design decision is well-motivated, as it is widely recognized that in word formation, in contrast to inflection, complex forms almost always have their idiosyncratic shades of meaning, even when classified as "transparent."

We have shown that the effect of morphological priming can be modeled precisely with a simple network that eschews morphemes and sublexical units, such as stems, affixes, and exponents, a result that is consistent with Word and Paradigm Morphology (Blevins, 2016) and the model of the mental lexicon proposed in Baayen et al. (2019). Do these findings imply that morphemes or morphs do not have any psychological or cognitive reality? Answering this question is not straightforward.

First, it is logically possible that morphemes are actually cognitively real and crucially involved in the lexical processing of German verbs. In this case, NDL is no more than a machine learning algorithm that generates correct predictions, but for the wrong reasons. To give substance to this argumentation, it will be

[13]When the analysis is restricted to the prime-target pairs in the Opaque and Transparent conditions, the model with NDL predictors again outperforms the model with classical predictors, now by 24.4 fREML units (for 8 edf, $p < 0.0001$). This provides further support against the present result being due to subjects using a task strategy using stem identity as response criterion.

[14]The learning rule of Rescorla and Wagner has several parameters that were introduced specifically to model differences in the salience of input features and the importance to the animal of different outcomes. In our implementation, we always set λ (representing the maximum amount of learning) to 1, and use a fixed learning rate (the product of the α and β parameters). In Baayen et al. (2011), the learning rate was 0.01, but subsequent work showed optimal performance when the learning rate is set to 0.001. For the simulation reported here, these values were used, and no simulations were run with different values.

[15]The NDL framework currently does not provide mechanisms that account for how the cognitive system reaches a response on the basis of lexome activations. Analyses with the generalized additive model indicate non-linear interactions for these mechanisms. At present, all we can do is bring these non-linearities to light with GAMs, in the hope that across experiments consistent patterns will emerge that then can be the subject of further computational modeling.

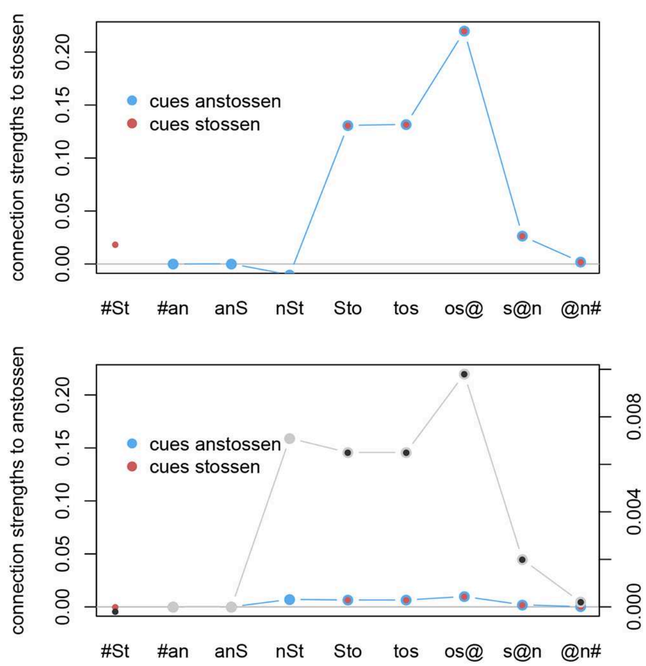


FIGURE 6 | Connection weights from triphones to the lexome STOSSEN (upper panel), and to the lexome ANSTOSSEN (lower panel). As the primes are of much lower frequency of occurrence than the targets, the magnitude of weights in the lower panel is substantially reduced compared to the upper panel. In the lower panel, the blue and red disks are presented on the same scale as the upper panel. The disks in gray and black represent the same datapoints, but now on an enlarged scale, depicted at the right-hand side of the plot.

essential for alternative explicit computational models to be put forward to champion the cause of units representing morphemes.

Second, the sublexical cues (triphones) that are shared by prime and target drive the prime-to-target pre-activation. For the prime-target pair ANSTOSSEN, STOSSEN, the set of shared trigrams is `Sto, tos, os@, s@n, s@n, @n#`, and the set of trigrams unique to only one of the two words is `#an, anS, nSt`, and `#St`. **Figure 6** presents the connection strengths of these cues to the lexome STOSSEN (upper panel) and ANSTOSSEN (lower panel). The cues that occur in both lexomes (the blue dots with inner red dots) support STOSSEN (upper panel), or ANSTOSSEN (lower panel) to exactly the same extent—as expected, as the cues are sublexical features that by definition must provide the same support for STOSSEN irrespective of whether they are embedded in the word form `Stos@n` or `anStos@n`.

Figure 6 brings to the fore two important points. First, the two central triphones of the stem that are shared by both the simple and the complex verb, `Sto` and `tos`, provide substantial contributions to the (pre-)activation of both STOSSEN and ANSTOSSEN, as expected. This observation fits well with the stem-frequency hypothesis, according to which the stem is the crucial unit mediating lexical access. Second, however, triphones at the boundary of the stem can have even greater strengths than these central triphones. For STOSSEN, this is the case for `os@`, and for ANSTOSSEN, this happens for both `nSt` and `os@`. The boundary triphone `s@n` also makes a non-negligible, albeit much smaller, contribution to the activation of these lexomes. Crucially, it is exactly here that NDL moves beyond the stem-frequency hypothesis. Triphones at the boundary of the stem often carry substantial discriminatory potential. The boundary triphone `os@` is important for discriminating (AN)STOSSEN from nominal lexomes, such as TROSS, STOSS, SCHLOSS and adjectival lexomes, such as GROSS, and the boundary triphone `nSt` is important for discriminating ANSTOSSEN from STOSSEN (note its high positive value as cue for ANSTOSSEN and its slightly negative value as cue for STOSSEN).

In other words, the present NDL model can be viewed as a refinement of the stem-frequency model that, by taking into account not only central sublexical features of the stem, but also the discriminatory potential of features at the stem boundary, achieves superior predictivity[16].

Third, NDL (and LDL) target implicit learning, the continuous recalibration of the lexicon that goes on without conscious thought and attention, similar to the way that object recognition is continually recalibrated (Marsolek, 2008). The finding of Smolka et al. (2015) that morphological priming effects are visible in evoked response potentials even in the absence of behavioral correlates is consistent with NDL capturing subliminal lexical processing. However, we also reflect on language, we enjoy word play, we have poetry, we teach grammar in schools, and in second language teaching we instruct learners how to put words together from their parts. Patients suffering from a stroke may benefit from explicit instruction about how inflected words can be put together (Nault, 2010). This knowledge about language is cognitively real, and it may also affect lexicality decisions in meta-linguistic tasks, such as primed lexicality decision making. NDL and LDL, however, are blind to this higher-order knowledge.

Fourth, in LDL, we can ask the question of how the triphone or trigram units used for its form vectors might be organized in two-dimensional space. The ordering of such units in the model's form vectors is not cognitively meaningful, but their organization in a 2-D plane, used as approximation of the cortical surface, might reveal interesting clustering. Depending on language and inflectional or derivational function, clusters in 2D space are indeed sometimes visible in triphone graphs when projected onto a plane, subject to constraints of stress minimization under self-organization (see Baayen et al., 2018, for detailed discussion). Topographic patterning under self-organization is also observed by Chersi et al. (2014), Marzi et al. (2018), who use temporal self-organizing maps (TSOMs). In their theory, trajectories in these maps capture, albeit fluidly, stems, and affixes. Since in this theory these trajectories subserve production, they can be viewed as morpheme-like units that are algorithmically functional for speech production. By contrast, the clusters of triphones that emerge under self-organization in NDL/LDL have no such algorithmically functional role. They typically do not form consecutive sequences of phones as found for stems and

[16]For detailed discussion of the importance of sublexical features at the boundary of morphemic units from a discrimination learning perspective, see Baayen et al. (2016b). For frailty at stem boundaries in speech production, see Baayen et al. (2019).

affixes in TSOMs. They emerge purely as a consequence of self-organizational constraints on spatial clustering of triphones given their overlap. If such clusters could be shown to correctly describe neuronal organization at more detailed levels of neurobiological modeling, then if such clusters were to be detected by brain imaging techniques, they should not be mistaken for evidence supporting algorithmically functional morphemes. In short, morpheme-like clusters can arise in NDL/LDL, without having any of the algorithmic functionalities commonly attributed to morphemes. Thus, these clusters of units are fundamentally different from neuronal clusters that might be hypothesized to underly localist morpheme units in interactive activation models.

Fifth, NDL and LDL are based on wide learning networks, simple two-layer networks with large numbers of units that can be trained very quickly, the mathematics of which are well-understood, and that can perform with surprising accuracy given well-chosen input and output representations (see e.g., Linke et al., 2017; Shafaei Bajestan and Baayen, 2018). Deep learning networks offer architectures in which units on hidden layers have the potential to become sensitive to, and in some sense "represent," morpheme-like units. Such networks are powerful statistical classifiers, but require decisions about the number of hidden layers, the number of units on these layers, and where to position convolutional and/or recurrent layers. Unfortunately, deep learning networks are widely recognized to have a "black box" nature, although progress is being made toward understanding why they work (see e.g., Cheng et al., 2018; Daniel and Yeung, 2019). NDL and LDL are specifically designed to provide both interpretational transparency and accurate and falsifiable predictions.

Importantly, what sets both wide learning and deep learning apart from the interactive activation framework is, first, that the former models are dynamically learning classifiers whereas the latter approach builds on the idea of a static classifier with a large number of parameters that have to be set manually, and second, that the former are end-to-end models whereas the latter solves only a partial task. The interactive activation framework is set up to select one word form and suppress all others, given visual input, but it remains silent about the semantics to which this form is supposed to provide access. This approach is still chained to the metaphor of the paper dictionary, in which form entries have to be located that, once found, provide access to meaning (see also Elman, 2009, for detailed criticism of the dictionary metaphor). By contrast, wide learning, following many practical applications in computational linguistics that make use of deep learning, is set up to predict the ultimate true goal of comprehension: the semantics targeted by the input signal. The results obtained with NDL and LDL obtained thus far suggest that this goal can be reached without mediation by form units, such as stems, affixes, or exponents. It is likely that future versions of the general LDL model of Baayen et al. (2019) will incorporate deep learning for some of its components, sacrificing interpretational transparency for increased accuracy. If morph-like units arise in such versions of the model, these units will not be part of a classical morphological calculus with symbolic representations and rules operating on these representations, such as proposed by Chomsky and Halle (1968) and Pinker (1999). Their function would be to statistically integrate high-dimensional evidence for semantics in interaction with large numbers of other such units.

In the light of these considerations, it is clear that the present study cannot provide a full answer to the question of whether morphemes are, or are not, cognitively real. Clearly, the behavioral findings that have been interpreted as evidence for stem-driven lexical access are real. What the present study adds to this is that there actually is another possible interpretation for the observed priming effects, namely, prime-to-target pre-activation in a discriminative lexicon. The discriminative lexicon provides detailed quantitative predictions within a larger conceptual framework that is informed by recent developments in linguistic morphology, using a computational algorithm that is completely driven by the distributional properties of the lexicon, and that does not require (nor allow) tuning of free parameters to bring the model's performance in line with the observed data. The present modeling results thus challenge proponents of morpheme-driven, decompositional, lexicons to demonstrate that their high-level conceptual theories can actually be made to work algorithmically. This in turn will make it possible to pit against each other the detailed quantitative predictions of morpheme-based and morpheme-free models.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

ES ran the experiment. RB did the modeling. Both the authors contributed to the writing of the manuscript.

FUNDING

This study was supported by the VolkswagenStiftung, Grant FP 561/11, awarded to ES, and by ERC advanced grant 742545-WIDE to RB.

ACKNOWLEDGMENTS

We acknowledge support from the Open Access Publishing Fund of University of Tübingen.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2020.00017/full#supplementary-material

REFERENCES

Arnold, D., Tomaschek, F., Lopez, F., Sering, T., and Baayen, R. H. (2017). Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit. *PLoS ONE* 12:e0174623. doi: 10.1371/journal.pone.0174623

Aronoff, M. (1994). *Morphology by Itself: Stems and Inflectional Classes*. Cambridge, MA: The MIT Press.

Baayen, R. H., Chuang, Y., and Blevins, J. P. (2018). Inflectional morphology with linear mappings. *Ment. Lexicon* 13, 232–270. doi: 10.1075/ml.18010.baa

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., and Blevins, J. (2019). The discriminative lexicon: a unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. *Complexity* 2019:4895891. doi: 10.1155/2019/4895891

Baayen, R. H., Dijkstra, T., and Schreuder, R. (1997). Singulars and plurals in Dutch: evidence for a parallel dual route model. *J. Mem. Lang.* 36, 94–117. doi: 10.1006/jmla.1997.2509

Baayen, R. H., McQueen, J., Dijkstra, T., and Schreuder, R. (2003). "Frequency effects in regular inflectional morphology: revisiting Dutch plurals," in *Morphological Structure in Language Processing*, eds R. H. Baayen and R. Schreuder (Berlin: Mouton de Gruyter), 355–390.

Baayen, R. H., Milin, P., Filipović Đurđević, D., Hendrix, P., and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychol. Rev.* 118, 438–482. doi: 10.1037/a0023851

Baayen, R. H., Milin, P., and Ramscar, M. (2016a). Frequency in lexical processing. *Aphasiology* 30, 1174–1220. doi: 10.1080/02687038.2016.1147767

Baayen, R. H., Piepenbrock, R., and van Rijn, H. (1993). *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Baayen, R. H., Shaoul, C., Willits, J., and Ramscar, M. (2016b). Comprehension without segmentation: A proof of concept with naive discriminative learning. *Lang. Cogn. Neurosci.* 31, 106–128. doi: 10.1080/23273798.2015.1065336

Baayen, R. H., Vasishth, S., Bates, D., and Kliegl, R. (2017). The cave of shadows. Addressing the human factor with generalized additive mixed models. *J. Mem. Lang.* 94, 206–234. doi: 10.1016/j.jml.2016.11.006

Bates, D. M., Kliegl, R., Vasishth, S., and Baayen, R. H. (2015). Parsimonious mixed models. *arxiv* 1506.04967.

Bauer, L., Lieber, R., and Plag, I. (2015). *The Oxford Reference Guide to English Morphology*. Oxford: Oxford University Press.

Beard, R. (1977). On the extent and nature of irregularity in the lexicon. *Lingua* 42, 305–341. doi: 10.1016/0024-3841(77)90102-4

Blevins, J. P. (2003). Stems and paradigms. *Language* 79, 737–767. doi: 10.1353/lan.2003.0206

Blevins, J. P. (2006). Word-based morphology. *J. Linguist.* 42, 531–573. doi: 10.1017/S0022226706004191

Blevins, J. P. (2016). *Word and Paradigm Morphology*. Oxford: Oxford University Press.

Booij, G. E. (1986). Form and meaning in morphology: the case of Dutch 'agent nouns'. *Linguistics* 24, 503–517. doi: 10.1515/ling.1986.24.3.503

Butz, M. V., and Kutter, E. F. (2016). *How the Mind Comes Into Being: Introducing Cognitive Science From a Functional and Computational Perspective*. Oxford: Oxford University Press.

Cheng, X., Khomtchouk, B., Matloff, N., and Mohanty, P. (2018). Polynomial regression as an alternative to neural nets. *arXiv* 1806.06850.

Chersi, F., Ferro, M., Pezzulo, G., and Pirrelli, V. (2014). Topological self-organization and prediction learning support both action and lexical chains in the brain. *Top. Cogn. Sci.* 6, 476–491. doi: 10.1111/tops.12094

Chomsky, N., and Halle, M. (1968). *The Sound Pattern of English*. New York, NY: Harper and Row.

Chuang, Y., Vollmer, M.-L., Shafaei-Bajestan, E., Gahl, S., Hendrix, P., and Baayen, R. H. (2019). *The Processing of Nonword Form and Meaning in Production and Comprehension: A Computational Modeling Approach Using Linear Discriminative Learning*. University of Tübingen.

Cieliebak, M., Dertu, J., Uzdilli, F., and Egger, D. (2017). "A twitter corpus and benchmark resources for german sentiment analysis," in *Proceedings of the 4th International Workshop on Natural Language Processing for Social Media (SocialNLP 2017)* (Valencia).

Creemers, A., Goodwin Davies, A., Wilder, R. J., Tamminga, M., and Embick, D. (2019). Opacity, transparency, and morphological priming: a study of prefixed verbs in Dutch. *J. Mem. Lang.* 110:104055. doi: 10.1016/j.jml.2019.104055

Daniel, W. B., and Yeung, E. (2019). A constructive approach for one-shot training of neural networks using hypercube-based topological coverings. *arXiv* 1901.02878.

Danks, D. (2003). Equilibria of the Rescorla-Wagner model. *J. Math. Psychol.* 47, 109–121. doi: 10.1016/S0022-2496(02)00016-0

De Grauwe, S., Lemhöfer, K., and Schriefers, H. (2019). Processing derived verbs: the role of motor-relatedness and type of morphological priming. *Lang. Cogn. Neurosci.* 34, 973–990. doi: 10.1080/23273798.2019.1599129

Deriu, J., Lucchi, A., De Luca, V., Severyn, A., Müller, S., Cieliebak, M., et al. (2017). "Leveraging large amounts of weakly supervised data for multi-language sentiment classification," in *Proceedings of the 26th International World Wide Web Conference (WWW-2017)* (Perth, WA). doi: 10.1145/3038912.3052611

Divjak, D., Milin, P., and Baayen, R. H. (2017). A learning perspective on individual differences in skilled reading: exploring and exploiting orthographic and semantic discrimination cues. *J. Exp. Psychol. Learn. Mem. Cogn.* 43, 1730–1751. doi: 10.1037/xlm0000410

Dunabeitia, J., Kinoshita, S., Carreiras, M., and Norris, D. (2011). Is morpho-orthographic decomposition purely orthographic? Evidence from masked priming in the same-different task. *Lang. Cogn. Process.* 26, 509–529. doi: 10.1080/01690965.2010.499215

Eisenberg, P. (2004). *Grundriss der deutschen Grammatik: Das Wort. 2nd Edn*. Vol. 1. Stuttgart; Weimar: J. B. Metzler.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., et al. (2012). A large-scale model of the functioning brain. *Science* 338, 1202–1205. doi: 10.1126/science.1225266

Elman, J. L. (2009). On the meaning of words and dinosaur bones: lexical knowledge without a lexicon. *Cogn. Sci.* 33, 547–582. doi: 10.1111/j.1551-6709.2009.01023.x

Fasiolo, M., Goude, Y., Nedellec, R., and Wood, S. (2017). *Fast Calibrated Additive Quantile Regression*. Available online at: https://github.com/mfasiolo/qgam

Feldman, L., and Prostko, B. (2001). Graded aspects of morphological processing: task and processing time. *Brain Lang.* 81, 1–16. doi: 10.1006/brln.2001.2503

Feldman, L. B., O'Connor, P. A., and Moscoso del Prado Martin, F. (2009). Early morphological processing is morpho-semantic and not simply morpho-orthographic: evidence from the masked priming paradigm. *Psychon. Bull. Rev.* 16, 684–691. doi: 10.3758/PBR.16.4.684

Feldman, L. B., Soltano, E. G., Pastizzo, M. J., and Francis, S. E. (2004). What do graded effects of semantic transparency reveal about morphological processing? *Brain Lang.* 90, 17–30. doi: 10.1016/S0093-934X(03)00416-4

Fleischer, W., and Barz, I. (1992). *Wortbildung der deutschen gegenwartssprache*. Tübingen: Max Niemeyer Verlag.

Frauenfelder, U. H., and Schreuder, R. (1992). "Constraining psycholinguistic models of morphological processing and representation: the role of productivity," in *Yearbook of Morphology 1991*, eds G. E. Booij and J. V. Marle (Dordrecht: Kluwer Academic Publishers), 165–183. doi: 10.1007/978-94-011-2516-1_10

Gonnerman, L., and Anderson, E. (2001). "Graded semantic and phonological similarity effects in morphologically complex words," in *Morphology 2000: Selected Papers From the 9th Morphology Meeting*, eds S. Bendjaballah, W. Dressler, O. E. Pfeiffer, and M. D. Voeikova (Amsterdam: John Benjamins), 137–148. doi: 10.1075/cilt.218.12gon

Gonnerman, L. M., Seidenberg, M. S., and Andersen, E. S. (2007). Graded semantic and phonological similarity effects in priming: evidence for a distributed connectionist approach to morphology. *J. Exp. Psychol. Gen.* 136:323. doi: 10.1037/0096-3445.136.2.323

Grainger, J., Dufau, S., Montant, M., Ziegler, J. C., and Fagot, J. (2012). Orthographic processing in baboons (*Papio papio*). *Science* 336, 245–248. doi: 10.1126/science.1218152

Günther, F., Smolka, E., and Marelli, M. (2019). 'Understanding' differs between English and German: capturing systematic language differences of complex words. *Cortex* 116, 168–175. doi: 10.1016/j.cortex.2018.09.007

Gurney, K., Prescott, T., and Redgrave, P. (2001). A computational model of action selection in the basal ganglia. *Biol. Cybern*. 84, 401–423. doi: 10.1007/PL00007984

Hannagan, T., Ziegler, J. C., Dufau, S., Fagot, J., and Grainger, J. (2014). Deep learning of orthographic representations in baboons. *PLoS ONE* 9:e84843. doi: 10.1371/journal.pone.0084843

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., et al. (2014). Deep speech: scaling up end-to-end speech recognition. *arXiv* 1412.5567.

Harm, M. W., and Seidenberg, M. S. (2004). Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychol. Rev*. 111, 662–720. doi: 10.1037/0033-295X.111.3.662

Herdağdelen, A., and Marelli, M. (2017). Social media and language processing: how facebook and twitter provide the best frequency estimates for studying word recognition. *Cogn. Sci*. 41, 976–995. doi: 10.1111/cogs.12392

Kappel, D., Habenschuss, S., Legenstein, R., and Maass, W. (2015). Network plasticity as bayesian inference. *PLoS Comput. Biol*. 11:e1004485. doi: 10.1371/journal.pcbi.1004485

Kappel, D., Legenstein, R., Habenschuss, S., Hsieh, M., and Maass, W. (2017). Reward-based stochastic self-configuration of neural circuits. *arXiv* 1704.04238.

Kielar, A., and Joanisse, M. F. (2011). The role of semantic and phonological factors in word recognition: an erp cross-modal priming study of derivational morphology. *Neuropsychologia* 49, 161–177. doi: 10.1016/j.neuropsychologia.2010.11.027

Kuperman, V., Bertram, R., and Baayen, R. H. (2008). Morphological dynamics in compound processing. *Lang. Cogn. Process*. 23, 1089–1132. doi: 10.1080/01690960802193688

Kuperman, V., Bertram, R., and Baayen, R. H. (2010). Processing trade-offs in the reading of Dutch derived words. *J. Mem. Lang*. 62, 83–97. doi: 10.1016/j.jml.2009.10.001

Kuperman, V., Schreuder, R., Bertram, R., and Baayen, R. H. (2009). Reading of multimorphemic Dutch compounds: towards a multiple route model of lexical processing. *J. Exp. Psychol*. 35, 876–895. doi: 10.1037/a0013484

Landauer, T., and Dumais, S. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychol. Rev*. 104, 211–240. doi: 10.1037/0033-295X.104.2.211

Lavric, A., Rastle, K., and Clapp, A. (2011). What do fully visible primes and brain potentials reveal about morphological decomposition? *Psychophysiology* 48, 676–686. doi: 10.1111/j.1469-8986.2010.01125.x

Lazaridou, A., Marelli, M., Zamparelli, R., and Baroni, M. (2013). "Compositionally derived representations of morphologically complex words in distributional semantics," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics - ACL (1)* (Sofia: Association for Computational Linguistics), 1517–1526.

Linke, M., Broeker, F., Ramscar, M., and Baayen, R. H. (2017). Are baboons learning "orthographic" representations? Probably not. *PLoS ONE* 12:e0183876. doi: 10.1371/journal.pone.0183876

Longtin, C., Segui, J., and Hallé, P. (2003). Morphological priming without morphological relationship. *Lang. Cogn. Process*. 18, 313–34. doi: 10.1080/01690960244000036

Lund, K., and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behav. Res. Methods Instr. Comput*. 28, 203–208. doi: 10.3758/BF03204766

Marelli, M., Amenta, S., Morone, E. A., and Crepaldi, D. (2013). Meaning is in the beholder's eye: morpho-semantic effects in masked priming. *Psychon. Bull. Rev*. 20, 534–541. doi: 10.3758/s13423-012-0363-2

Marelli, M., and Baroni, M. (2015). Affixation in semantic space: modeling morpheme meanings with compositional distributional semantics. *Psychol. Rev*. 122:485. doi: 10.1037/a0039267

Marslen-Wilson, W., Older, L., Janssen, D., and Waksler, S. (1994a). "Stems and affixes: local and non-local attachment," in *Paper Presented to the Experimental Psychology Society* (London).

Marslen-Wilson, W., Tyler, L. K., Waksler, R., and Older, L. (1994b). Morphology and meaning in the English mental lexicon. *Psychol. Rev*. 101, 3–33. doi: 10.1037/0033-295X.101.1.3

Marsolek, C. J. (2008). What antipriming reveals about priming. *Trends Cogn. Sci*. 12, 176–181. doi: 10.1016/j.tics.2008.02.005

Marzi, C., Ferro, M., and Pirrelli, V. (2018). "Is inflectional irregularity dysfunctional to human processing?" in *Abstract Booklet, The Mental Lexicon 2018*, ed V. Kuperman (Edmonton, AB: University of Alberta), 60.

Matthews, P. H. (1974). *Morphology. An Introduction to the Theory of Word Structure*. Cambridge: Cambridge University Press.

McClelland, J. L., and Rumelhart, D. E. (Eds.). (1986). *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. 2: Psychological and Biological Models*. Cambridge, MA: MIT Press.

Meunier, F., and Longtin, C. (2007). Morphological decomposition and semantic integration in word processing. *J. Mem. Lang*. 56, 457–471. doi: 10.1016/j.jml.2006.11.005

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, eds C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Curran Associates, Inc), 26, 3111–3119. Available online at: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., and Baayen, R. H. (2017a). Discrimination in lexical decision. *PLoS ONE* 12:e0171935. doi: 10.1371/journal.pone.0171935

Milin, P., Smolka, E., and Feldman, L. B. (2017b). "Models 11 of lexical access and morphological processing," in *The Handbook of Psycholinguistics*, eds E. M. Fernández and H. S. Cairns (Malden: Wiley-Blackwell), 240–268. doi: 10.1002/9781118829516.ch11

Nault, K. (2010). *Morphological therapy protocol*. (Ph. D. thesis), University of Alberta, Edmonton, AB.

Napps, S. E., and Fowler, C. A. (1987). Formal relationships among words and the organization of the mental lexicon. *J Psycholinguist Res.* 16, 257–272.

Norris, D., and Kinoshita, S. (2008). Perception as evidence accumulation and bayesian inference: insights from masked priming. *J. Exp. Psychol*. 137, 434–455. doi: 10.1037/a0012799

Pastizzo, M. J., and Feldman, L. B. (2002). Does prime modality influence morphological processing? *Brian Lang*. 81, 28–41. doi: 10.1006/brln.2001.2504

Pinker, S. (1999). *Words and Rules: The Ingredients of Language*. London: Weidenfeld and Nicolson.

Plag, I. (2003). *Word-Formation in English*. Cambridge: Cambridge University Press.

Plaut, D. C., and Gonnerman, L. M. (2000). Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Lang. Cogn. Process*. 15, 445–485. doi: 10.1080/01690960050119661

Rastle, K., and Davis, M. (2008). Morphological decomposition based on the analysis of orthography. *Lang. Cogn. Process*. 23, 942–971. doi: 10.1080/01690960802069730

Rastle, K., Davis, M. H., Marslen-Wilson, W., and Tyler, L. K. (2000). Morphological and semantic effects in visual word recognition: a time-course study. *Lang. Cogn. Process*. 15, 507–537. doi: 10.1080/01690960050119689

Rastle, K., Davis, M. H., and New, B. (2004). The broth in my brother's brothel: morpho-orthographic segmentation in visual word recognition. *Psychon. Bull. Rev*. 11, 1090–1098. doi: 10.3758/BF03196742

Redgrave, P., Prescott, T., and Gurney, K. (1999). The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience* 89, 1009–1023. doi: 10.1016/S0306-4522(98)00319-4

Rescorla, R. A., and Wagner, A. R. (1972). "A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement," in *Classical Conditioning II: Current Research and Theory*, eds A. H. Black and W. F. Prokasy (New York, NY: Appleton Century Crofts), 64–99.

Roelofs, A. (1997a). A case for nondecomposition in conceptually driven word retrieval. *J. Psycholinguist. Res*. 26, 33–67. doi: 10.1023/A:1025060104569

Roelofs, A. (1997b). "Morpheme frequency in speech production: testing WEAVER," in *Yearbook of Morphology 1996*, eds G. E. Booij and J. Van Marle (Dordrecht: Kluwer), 135–154.

Roelofs, A., Baayen, R. H., and van den Brink, D. (2002). Morphology by itself in planning the production of spoken words. *Psychon. Bull. Rev*. 9, 132–138. doi: 10.3758/BF03196269

Rumelhart, D. E., and McClelland, J. L. (1986). "On learning the past tenses of English verbs," in *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. 2: Psychological and Biological Models*, eds J. L. McClelland and D. E. Rumelhart (Cambridge, MA: The MIT Press), 216–271.

Schmidtke, D., Matsuki, K., and Kuperman, V. (2017). Surviving blind decomposition: a distributional analysis of the time-course of complex word recognition. *J. Exp. Psychol. Learn. Mem. Cogn*. 43:1793. doi: 10.1037/xlm0000411

Schreuder, R. (1990). Lexical processing of verbs with separable particles. *Yearb. Morphol*. 3, 65–79.

Schreuder, R., Burani, C., and Baayen, R. H. (2003). "Parsing and semantic opacity" in *Reading Complex Words. Cross-Language Studies*, eds E. Assink and D. Sandra (Dordrecht: Kluwer), 159–189. doi: 10.1007/978-1-4757-3720-2_8

Schreuder, R., Grendel, M., Poulisse, N., Roelofs, A., and Van de Voort, M. (1990). "Lexical processing, morphological complexity and reading," in *Comprehension Processes in Reading*, eds D. A. Balotta, G. B. Flores d'Arcais, and K. Rayner (Hillsdale, NJ: Erlbaum), 125–141.

Schriefers, H., Zwitserlood, P., and Roelofs, A. (1991). The identification of morphological complex spoken words: continuous processing or decomposition? *J. Mem. Lang*. 30, 26–47. doi: 10.1016/0749-596X(91)90009-9

Sering, K., Weitz, M., Kuenstle, D., and Schneider, L. (2017). *Pyndl: Naive Discriminative Learning in Python*. doi: 10.5281/zenodo.597964

Sering, T., Milin, P., and Baayen, R. H. (2018b). Language comprehension as a multiple label classification problem. *Stat. Neerland*. 72, 339–353. doi: 10.1111/stan.12134

Shafaei Bajestan, E., and Baayen, R. H. (2018). "Wide learning for auditory comprehension," in *Interspeech* (Hyderabad). doi: 10.21437/Interspeech.2018-2420

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature* 529:484. doi: 10.1038/nature16961

Smolka, E., and Eulitz, C. (2011). "Asymmetric meaning assembly for semantically transparent and opaque complex verbs in German," in *Paper Presented at the Seventh International Morphological Processing Conference (MOPROC)* (San Sebastian).

Smolka, E., and Eulitz, C. (2018). Psycholinguistic measures for German verb pairs: semantic trans-parency, semantic relatedness, verb family size, and age of reading acquisition. *Behav. Res. Methods* 50, 1540–1562. doi: 10.3758/s13428-018-1052-5

Smolka, E., Gondan, M., and Rösler, F. (2015). Take a stand on understanding: electrophysiological evidence for stem access in german complex verbs. *Front. Hum. Neurosci*. 9:62. doi: 10.3389/fnhum.2015.00062

Smolka, E., Komlosi, S., and Rösler, F. (2009). When semantics means less than morphology: the processing of German prefixed verbs. *Lang. Cogn. Process*. 24, 337–375. doi: 10.1080/01690960802075497

Smolka, E., Libben, G., and Dressler, W. U. (2019). When morphological structure overrides meaning: evidence from German prefix and particle verbs. *Lang. Cogn. Neurosci*. 34, 599–614. doi: 10.1080/23273798.2018.1552006

Smolka, E., Preller, K. H., and Eulitz, C. (2014). 'verstehen' ('understand") primes 'stehen' ('stand'): morphological structure overrides semantic compositionality in the lexical representation of German complex verbs. *J. Mem. Lang*. 72, 16–36. doi: 10.1016/j.jml.2013.12.002

Smolka, E., Zwitserlood, P., and Rösler, F. (2007). Stem access in regular and irregular inflection: evidence from German participles. *J. Mem. Lang*. 57, 325–347. doi: 10.1016/j.jml.2007.04.005

Stewart, T. C., Bekolay, T., and Eliasmith, C. (2012). Learning to select actions with spiking neurons in the basal ganglia. *Front. Neurosci*. 6:2. doi: 10.3389/fnins.2012.00002

Stump, G. (2001). *Inflectional Morphology: A Theory of Paradigm Structure*. Cambridge: Cambridge University Press.

Taft, M., and Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *J. Verb. Learn. Verb. Behav*. 14, 638–647. doi: 10.1016/S0022-5371(75)80051-X

Tomaschek, F., Plag, I., Ernestus, M., and Baayen, R. H. (2019). Phonetic effects of morphology and context: Modeling the duration of word-final S in english with naive discriminative learning. *J. Linguist*. 1–39. doi: 10.31234/osf.io/4bmwg

Tomaschek, F., Tucker, B. V., Fasiolo, M., and Baayen, R. H. (2018). Practice makes perfect: the consequences of lexical proficiency for articulation. *Linguist. Vanguard* Band 4: Heft s2:20170018. doi: 10.1515/lingvan-2017-0018

van Rij, J., Wieling, M., Baayen, R. H., and van Rijn, H. (2017). *itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs*. R package version 2.3.

Veríssimo, J. (2018). "Taking it a level higher: the leia model of complex word recognition," in *Poster Presented at AMLaP 2018* (Berlin).

Wood, S. N. (2017). *Generalized Additive Models*. New York, NY: Chapman & Hall; CRC. doi: 10.1201/9781315370279

Zwitserlood, P. (2018). "Processing and representation of morphological complexity in native language comprehension and production," in *The Construction of Words. Studies in Morphology*, ed G. E. Booij (Cham: Springer), 4, 583–602.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Communication

ORIGINAL RESEARCH
published: 22 December 2020
doi: 10.3389/fcomm.2020.434106

Check for updates

Revisiting Aspect in Mild Cognitive Impairment and Alzheimer's Disease: Evidence From Greek

Christina Manouilidou[1]*, *Georgia Roumpea*[1], *Anastasia Nousia*[2], *Stavroula Stavrakaki*[3] *and Grigorios Nasios*[2]

[1] Department of Comparative and General Linguistics, University of Ljubljana, Ljubljana, Slovenia, [2] Department of Speech and Language Therapy, School of Health Sciences, University of Ioannina, Ioannina, Greece, [3] Department of Italian Language and Literature, Aristotle University, Thessaloniki, Greece

The study investigates the ability of Greek-speaking individuals diagnosed with mild Alzheimer's Disease (mAD) and Mild Cognitive Impairment (MCI) to produce verbs that vary with respect to their grammatical and lexical aspect. While grammatical aspect has been examined in aphasia, there are only a few studies dealing with this in neurodegenerative conditions and their findings are contradictory. Motivated by this, we further investigate aspect by examining not only grammatical but lexical aspect as well and how their semantic and temporal features affect mAD and MCI individuals' performance. Thus, the major innovation of the study is that it examines aspect not only as a functional feature but also as a lexical variable, something addressed for the first time in the literature. We also address whether grammatical aspect interacts with lexical aspect and with time reference. Finally, by looking at Greek, we further contribute to cross-linguistic perspective of aspect investigation. 11 MCI and 11 mAD individuals participated in a picture naming task, targeting the investigation of lexical aspect, and a sentence completion task, targeting the investigation of grammatical aspect and its interaction with lexical aspect and time reference. Both groups of participants were found to be impaired in both tasks when compared to healthy controls. In the naming task, both group and lexical aspect were significant predictors for participants' performance. Specifically, more impaired performance was found in *states* (*believe*), *achievements* (*break*), and *semelfactives* (*hit*) compared to *activities* (*run*) and *accomplishments* (*build*) for both AD and MCI participants. In the sentence completion task, apart from group, neither grammatical or lexical aspect nor tense were significant predictors for participants' performance. While results indicate that both grammatical and lexical aspect are impaired in AD and MCI, a closer look suggests a dissociation regarding the temporal feature of *duration*. Specifically, as grammatical feature, *duration* does not appear to affect participants' choice between perfective and imperfective aspect. As a lexical variable, on the other hand, and as part of the lexical representation of a verb, duration (together with internal structure) appears to play a role in verb naming. Finally, the lack of interaction between lexical and grammatical aspect also indicates that these two subsystems can be affected differentially.

Keywords: sentence completion task, picture naming task, grammatical aspect, lexical aspect, Alzheimer's disease, mild cognitive impairment, Greek language

OPEN ACCESS

Edited by:
Gonia Jarema,
Université de Montréal, Canada

Reviewed by:
Valantis Fyndanis,
Cyprus University of Technology, Cyprus
Gary Libben,
Brock University, Canada

Correspondence:
Christina Manouilidou
christina.manouilidou@ff.uni-lj.si

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *31 October 2018*
Accepted: *25 November 2020*
Published: *22 December 2020*

Citation:
Manouilidou C, Roumpea G, Nousia A, Stavrakaki S and Nasios G (2020) Revisiting Aspect in Mild Cognitive Impairment and Alzheimer's Disease: Evidence From Greek. Front. Commun. 5:434106. doi: 10.3389/fcomm.2020.434106

INTRODUCTION

Alzheimer's disease (AD) is a chronic neurodegenerative disease and the most common type of dementia (Visser et al., 1999), characterized by progressive cognitive dysfunction. Diagnosing dementia based on neuropsychological assessments requires the presence of impairment in the domain of memory and in one of the other cognitive domains (Lindeboom and Weinstein, 2004). At the initial stages of AD, working memory is impaired (Braaten et al., 2006), leading to difficulties in learning new things. As the disease progresses, dysfunction in other cognitive domains, such as executive functions, attention and visuospatial skills, is observed. The term Mild Cognitive Impairment (MCI), refers to the clinical stage between normal aging and dementia. Individuals diagnosed with MCI suffer from loss of cognitive and functional abilities, yet they do not meet the criteria to be diagnosed with dementia (Petersen et al., 2001). Patients who demonstrate impairment in multiple cognitive domains, with or without degraded memory, including language, are more likely to develop dementia (Petersen et al., 2001; Petersen, 2003; Alexopoulos et al., 2006).

Language abilities in AD are affected during all stages of the disease with patients having difficulties in both production and comprehension of both grammatical and semantic aspects of language (Altmann et al., 2001; Kavé and Dassa, 2018; Kavé and Goral, 2018). An early symptom of AD is difficulties in producing and recalling single words due to impairment at different levels of lexical processing. Recent studies employing a revised version of Boston Naming Test and a picture-naming task (Silagi et al., 2015; Salehi et al., 2017) report noun naming deficits in AD patients. These studies revealed that naming errors are different in terms of quantity and quality among the different stages of the disease (mild and moderate), with patients at the initial stage producing more semantic errors and patients at the moderate stage producing fewer correct responses and mainly no-responses at all. According to the authors (Silagi et al., 2015; Salehi et al., 2017), patients' lexical retrieval difficulties might indicate either deficit in accessing the semantic content of a word or difficulties in recalling its phonological form. While nouns are, in general, more impaired than verbs (Whatmough and Chertkow, 2002), Masterson et al. (2007) observed more errors and slower reaction times to verbs compared to nouns in a picture-naming task and a word-picture verification task. The above findings are in line with other studies which reveal that verb production and comprehension abilities are more impaired than noun naming abilities in AD (Kim and Thompson, 2004; Druks et al., 2006). Finally, verb impairment in AD has also been observed in verbal fluency tasks, where patients demonstrated recall and production difficulties (Alegret et al., 2018). In general, it appears that impaired naming abilities of individuals with AD result from patients' general cognitive impairment and specifically from their manifested memory limitations. Specifically, it has been suggested that degraded semantic memory, the part of long-term memory which includes language and mental lexicon information, might interfere with patients' naming abilities (Vogel et al., 2005; Braaten et al., 2006).

Studies investigating morphosyntactic abilities in AD have produced mixed results. For instance, Kavé and Levy (2003) found AD patients' speech to be less informative with more semantic errors compared to controls, but patients' syntactic (e.g., production of independent and declarative clauses) and morphological (e.g., word inflection, verb formation) abilities were comparable to the abilities of the cognitively intact participants. However, other studies reveal general morphosyntactic impairment in AD, such as impaired verb morphology (Walenski et al., 2009), morphosyntactic errors (e.g., incorrect inflections, word order errors, and missing matrix or subordinate clauses) in spontaneous speech and oral production (Altmann et al., 2001), and difficulties in interpreting thematic roles of verbs (e.g., Manouilidou and de Almeida, 2009; Manouilidou et al., 2009). Mixed results have also been reported for functional categories associated with verbs, like tense, aspect, agreement and mood. Some studies report unimpaired performance in tense and agreement (Appell et al., 1982; Bayles, 1982; Kaprinis and Stavrakaki, 2007) while others describe difficulties in production of tense, aspect, agreement and mood (Fyndanis et al., 2013, 2018; Roumpea et al., 2019).

Fyndanis et al. (2013) found Greek-speaking mAD individuals to be impaired in production and comprehension of agreement, tense and grammatical aspect in a sentence-completion task and a grammaticality judgment task. Grammatical aspect was found to be more impaired than tense and agreement both in production and grammaticality judgment. The authors attributed this finding to the higher processing demands of grammatical aspect which requires the integration of both grammatical and conceptual information (following Kok et al., 2007). Moreover, a difference between perfective (*I broke*) and imperfective (*I was breaking*) emerged, with imperfective being significantly more impaired contra to the suggestion that unmarked values are better preserved than marked values (Lapointe, 1985). In a pilot study, Roumpea et al. (2019) examined perfective and imperfective grammatical aspect in MCI and AD. The authors predicted that the different features of perfective and imperfective (see section Grammatical Aspect) would affect participants' choice leading to a better performance on perfective given that its semantic and temporal features might make it less complex and easier to process compared to the imperfective. More specifically, imperfective aspect (*I was building*) presents the situation as an event that lasts in time and consists of different phases. In other words, it encodes an event that has duration as well as internal structure. This means that whenever AD and MCI individuals want to present an event in an imperfective way, they first have to think and create in their minds a continuous process with all its stages (beginning, middle, and end) which is probably highly demanding and difficult for them. On the other hand, perfective aspect (*I built*) presents a situation as a whole, complete event with no duration and no internal structure. Thus, whenever AD and MCI individuals present an event in a perfective way, they have to create in their minds a process that has already been completed with no internal structure, which probably makes it less complex and less demanding compared to imperfective. Thus, it can result in better performance. Three MCI individuals and one mild AD individual were tested in a sentence completion task tapping into production of grammatical aspect. Both MCI and mAD participants were found to be significantly impaired compared to controls. However, results reported no significant preference of perfective (*I broke*) over imperfective (*I was breaking*) aspect (contra Fyndanis et al., 2013). These findings suggest that grammatical aspect as a functional category is degraded, but the different semantic and temporal features of perfective (no duration, no internal structure) and imperfective (duration and internal structure) aspect do not seem to affect participants' performance.

With respect to language impairment in individuals with MCI, there exists plentiful evidence from standardized tests (for a review, see Taler and Phillips, 2008) but not much from psycholinguistic studies. Concerning word finding abilities and verbal fluency, results are controversial. Some studies report no impairment (e.g., Albert et al., 2007), while others found word generation and retrieval process to be compromised in both phonemic and category verbal fluency tasks (Demetriou and Holtzer, 2017). MCI patients were, also, found to have difficulties in recalling and producing verbs in a verb fluency task (Alegret et al., 2018). Concerning morphological knowledge and syntactic structure, studies reveal controversial findings with MCI individuals either being impaired (Lambon Ralph et al., 2003) or performing equally well with the control group (e.g., de Jager et al., 2003). In a recent study, Manouilidou et al. (2016) examined MCI individuals' abilities to detect morphological violations in an off-line grammaticality judgment task and an on-line lexical-decision task. Results revealed that patients' structural knowledge was not affected but processing morphological structure was impaired especially in the lexical-decision task due to time pressure. Also, functional categories, such as grammatical aspect, have also been found impaired (Roumpea et al., 2019).

MCI individuals' language difficulties have been attributed to impairments in episodic, working (Summers and Saunders, 2012) and semantic memory (Wilson et al., 2011), processing speed limitations, impaired attention and executive dysfunction (Summers and Saunders, 2012). Duong et al. (2006), by employing the Stroop picture naming task, suggested that MCI patients' performance might be affected by the type of task they are asked to perform and not only by their language abilities. Increased task complexity might lead MCI patients to a low performance, indicating that impaired executive functions can also interfere with language processing. In a similar vein, Manouilidou et al. (2016) found strong correlations between executive dysfunction in MCI and impaired language performance.

Language impairment in functional categories, such as tense and grammatical aspect, has been the focus of studies dealing not only with MCI or AD, but with aphasic populations too. Dragoy and Bastiaanse (2013) examined tense reference (past, present and future) and aspect (perfective and imperfective) in the Russian version of the Test of Assessment of Time reference. They report that non-fluent and fluent aphasic Russian-speakers are better on producing past-reference, especially in perfective context (compared to imperfective), and non-past reference in imperfective context (compared to perfective). The authors explained these findings in terms of prototypical matches between tense and aspect semantics. Imperfective verbs are prototypically used to refer to on-going events, thus, non-past tenses (present, future), while perfective ones prototypically describe completed events, thus past tenses. Also, Koukoulioti (2013) examined aspect in Greek-speaking individuals with semantic dementia and observed an interaction with verbal telicity. That is, in present tense (always imperfective in Greek), unaccusative verbs (telic) yielded worse performance than unergatives (atelic), a difference which was not observed in the past tense. Dragoy and Bastiaanse's (2013) findings were not supported by Fyndanis and Themistocleous (2019) who provided evidence from Greek-speaking aphasic and healthy participants. Tense and aspect reference were examined in a sentence-completion task. Both aphasic and healthy participants performed comparably on past and future reference independently of aspectual context. Furthermore, no dissociation between perfective and imperfective aspect was found and no interaction emerged between aspect and time reference in both aphasic and healthy participants.

Taking into account the few studies on functional and lexical categories associated with verbs, as well as the existing contradictory results, the present research aims to further investigate Greek-speaking mAD and MCI individuals' ability to produce verbs that vary with respect to their grammatical and lexical aspect. In particular, we investigated how the different temporal and semantic features of these categories would affect patients' language performance. An interaction between grammatical and lexical aspect was expected to emerge, given that the temporal properties of lexical and grammatical aspect not always overlap. For instance, non-durative verbs, like achievements, are more naturally expressed in perfective aspect while durative verbs are more naturally expressed in imperfective aspect (see section Lexical Aspect). To our knowledge this study is the first attempt to examine whether these two types of aspect interact and how this could interfere with language processing in Greek pathological populations. Finally, we also examined our results with respect to time reference and its interaction with grammatical aspect, an issue mainly investigated in aphasic populations.

LINGUISTIC BACKGROUND

Grammatical Aspect

Grammatical aspect is considered to be a functional category (Chomsky, 1995, 2000, 2001) which conveys information about time and it is often confused with tense. In fact, tense refers to *when* a situation takes place and relates it usually with the moment of speaking (past, present, future), while aspect provides information about *how* a situation takes place, or in other words about the internal temporal constituency of a situation (Comrie, 1976).

In Greek, like in other languages, grammatical aspect is overtly marked on the verb (Comrie, 1976; Holton et al., 2010; Moser, 2013). Consider, for example, the singing event in (1a) and (1b). Although both sentences refer to the past, describing an event prior to the moment of speaking, they differ regarding

TABLE 1 | Time reference and aspect of the verb *pezo* "I play" in Greek, with the imperfective stem *pez-* and the perfective stem *peks-*.

	Perfective	Imperfective
Present	N.A.	péz-o "I play"
Past	é-peks-a "I played"	é-pez-a "I was playing"
Future	θa péks-o "I will play"	θa péz -o "I will be playing"

how the event pertains to the past (progressively or non-progressively). This internal difference refers to grammatical aspect and more specifically to the aspectual distinction, that is, the distinction between imperfective (1a, *traγuðúsa* "I was singing" and perfective (1b, *traγúðisa* "I sang") These two grammatical aspect values have different semantic features. Perfective aspect describes a situation-action as a whole and complete event with no duration and pays no attention to its internal phases (Comrie, 1976; Holton et al., 2010; Moser, 2013). In contrast, imperfective aspect refers to the situation as an event with duration and internal phases (beginning, middle, and end).

(1a) *Xθes, eγó traγuðúsa, ótan to korítsi χtípise to kuðúni.*
"Yesterday, I was singing when the girl rang the bell"
(1b) *Xθes, eγó traγúðisa ðío traγúðja.*
"Yesterday, I sang two songs"

In Greek, time reference interacts with grammatical aspect. Perfective and imperfective are distinguished in past and future tenses, while the present tense morphologically encodes only imperfective aspect as it usually refers to a situation happening simultaneously with the moment of speaking and as such it cannot form perfective aspect (Comrie, 1976; Holton et al., 2010; Moser, 2013). **Table 1** illustrates the interaction of time reference and aspect in Greek.

Lexical Aspect

Lexical aspect is a semantic category inherent to the verb (Comrie, 1976; Smith, 1997; Moser, 2013). Verbs are divided into five categories with different semantic and temporal features: activities (*run*), accomplishments (*build*), semelfactives (*hit*), achievements (*break*) and states (*know*) (Smith, 1997).

Activity and accomplishment verbs share some temporal features, but they differ with respect to their end point. Both activities and accomplishments are durative, describing processes which last in time and, dynamic, which means that they are subject to an input of motion and involve change (Comrie, 1976). These verbs have internal structure as they are not homogeneous processes (Comrie, 1976; Smith, 1997). Consider a *"running"* (activity) event and a *"building"* (accomplishment) event. In both processes there is a necessary change of state as they evolve. When someone runs, there are moments that one foot is on the ground, the other one is not and so on (Comrie, 1976). Similarly, during the process of *"building"* there are different successive phases in which the process advances to its end point. The difference between activities and accomplishments is on the feature of telicity (Smith, 1997). Activities do not have a natural end point in which they complete (atelic), while accomplishments are processes with a natural final endpoint (telic) and when they reach this outcome, they are complete and no longer continue.

Concerning semelfactive and achievement verbs, they are dynamic and instantaneous, describing single-stage events, which occur very quickly and cannot be associated with the notion of duration (Smith, 1997). Examining the events of *"coughing," "I coughed"* (semelfactive), and *"breaking," "I broke the vase"* (achievement), we observe that *"coughing"* has no result or outcome, while *"breaking"* leads to a change of state, which is not, however, part of the event. Thus, semelfactives are atelic, while achievements are telic. Also, both semelfactives and achievements lack internal structure as they refer to dynamic situations as a single complete whole (Comrie, 1976). State verbs, on the other hand, are static and durative, and as such they refer to situations that are stable and last either for a moment or an interval (Smith, 1997). State verbs still retain the property of duration, even in cases that they last for a moment. Consider the sentence "*The temperature is* ninety *and rising*" [as mentioned in Smith (1997)], where the state *(temperature is ninety)* has short duration. This sentence, according to Smith (1997), is true as there is a minimal period where the state *(temperature is ninety)* remains and then there has to be a change of state *(temperature rising).* Finally, state verbs do not have internal structure, as they consist of undifferentiated stages (Comrie, 1976; Smith, 1997). For example, in the sentence *"I know how to write,"* all phases of *"know,"* whenever we choose to examine them, are going to be identical.

Based on their temporal features, verbs appear more commonly either in their perfective or in their imperfective form. Instantaneous verbs which do not last in time do not commonly appear in their imperfective form (Comrie, 1976; Moser, 2013). Thus, achievement and semelfactive verbs might appear more in their perfective aspect as instantaneous verbs without duration. As far as state verbs are concerned, they are likely to appear more in the imperfective aspect as they describe a stable situation, without alternations in time (Moser, 2013). Similarly, activity and accomplishment verbs might appear more in the imperfective aspect as verbs describing processes with duration (Comrie, 1976). **Table 2** summarizes the main features of each verb category by using examples from Greek.

RESEARCH QUESTIONS AND PREDICTIONS OF THE CURRENT STUDY

In the current study, we seek new evidence regarding the production and comprehension of grammatical and lexical aspect in mAD and MCI and how the two might interact or how performance on lexical aspect might significantly predict performance on grammatical aspect. In this section, we outline our predictions with a particular focus on the role of duration which is a feature found in both lexical and grammatical aspect.

Our prediction about *grammatical* aspect is that it will be impaired in both MCI and mAD participants, similarly to our previous study (Roumpea et al., 2019). With respect to the distinction between perfective and imperfective, if indeed

TABLE 2 | Summary of verb categories in lexical aspect together with their temporal features and examples in Greek.

Lexical aspect	Durative	Dynamic	Instantaneous	Stable	Telic	Atelic	Verb example
Activity	+	+	−	−	−	+	tréχo "run"
Accomplishment	+	+	−	−	+	−	χtízo "build"
Semelfactive	−	+	+	−	−	+	χtipáo "hit"
State	+	−	−	+	−	−	kséro "know"
Achievement	−	+	+	−	+	−	spáo "break"

duration is a decisive factor when processing aspect, then we expect participants to perform better on perfective (*I broke*) than imperfective aspect (*I was breaking*), even though the former is the marked value in Greek. Finally, it seems possible that grammatical aspect will interact with time reference, that is we hypothesize that tense will predict grammatical aspect, leading participants to perform better on perfective aspect within a past context and better on imperfective aspect within a future context. While this assumption is supported by evidence from Russian-speaking individuals with aphasia (Dragoy and Bastiaanse, 2013) but not by evidence from Greek-speaking individuals with aphasia (Fyndanis and Themistocleous, 2019) there is no evidence coming from any language about AD and MCI individuals, highlighting the contribution of the current study.

When it comes to *lexical aspect,* impaired recall and verb naming is expected in general, based on evidence provided by previous studies (Druks et al., 2006; Alegret et al., 2018). With respect to the different verb categories (activities, accomplishments, states, semelfactives, and achievements) our predictions are based on what we know about grammatical aspect, given that there are no studies dealing with lexical aspect in neurodegenerative conditions. Therefore, given that imperfectivity was found to pose more difficulties for participants, we can also assume that the feature of duration, as a lexical variable, will interfere with participants' performance. In other words, we would expect lower performance in inherently durative verbs. Thus, achievements (*break*) and semelfactives (*hit*) are expected to be better preserved as they describe instantaneous events, which lack duration and internal structure. On the other hand, activities (*run*), states (*know*), and accomplishments (*build*) are temporally more complex as they describe processes or events with duration and internal structure. They can be perceived as on-going and continuous events and as such they might be more demanding in terms of processing for MCI and mAD participants compared to instantaneous verbs.

Of particular interest is the possibility of an interaction between grammatical and lexical aspect given that they both convey temporal information which might not always be on a par. Namely, participants might have difficulties in attributing durative meanings to instantaneous verbs (achievement "*break,*" semelfactive "*hit*"), in other words, in selecting imperfective aspect for these verb categories. In contrast, perfective forms could be favored even in contexts where the right form is the imperfective. Similarly, although activity (*run*), state (*know*) and accomplishment (*build*) verbs, appear mainly in imperfective forms (Comrie, 1976), a better performance on the perfective might be expected, if processing duration poses difficulties.

Before launching into a description of the current investigation, it is important to situate it within the "Words in the World" scope and discuss the essential role *aspect* plays in communication. As situations unfold in time, their accurate temporal description plays a key role in our understanding. There are many ways languages of the world convey such temporal information but it is mostly through tense which specifies the location of an event in time. Whether an event has taken place in the past or will take place in the future constitutes an objective, undeniable fact which leaves each speaker with no personal choice. The correct choice of tense guarantees accuracy in communication.

Apart from accuracy, languages also express subjectivity through temporal expressions and this is mainly done through *aspect*. Aspectual information encodes the viewpoint of the speaker on a particular situation. It reflects the speaker's subjective choice to see an event as a whole, in its totality (by choosing perfectivity) or as it unfolds in time (by choosing imperfectivity). The event remains the same but what aspect does is that it adds an additional dimension to it which is bound by the speaker's perspective. Even though this is extralinguistic, subjective information, it is encoded in languages of the world in a variety of ways. Thus, expressing *aspect* requires an extra effort of combining linguistic with extra-linguistic information, a process which becomes fairly demanding for people with cognitive decline.

The interaction of the two, *tense* and *aspect* has already been examined in pathological populations (see references above). What has not been examined yet is their interaction with *lexical* aspect as well. Lexical aspect of a verb provides a sort of guide as to how to regulate tense and grammatical aspect. That is, imperfective grammatical aspect takes an internal view of an event and as such it is compatible with durative predicates of activities and accomplishments as it is congruent with progressive meaning. Similarly, perfective aspect is compatible with achievements given that their inherent lack of duration. Having said that, languages do allow combinations of perfective aspect and durative verbs (*she built a house*) as well as combinations of imperfective aspect and non-durative predicates (*as she was reaching the peak*). Such cases of incompatibility can be proven especially challenging for populations with cognitive decline, such as MCI and AD. The source of this incompatibility is not linguistic in nature, but it is very often encoded in languages.

The challenge of all this lies in the combination of linguistic with extra-linguistic information which heavily affects the processing of these forms by patients with cognitive decline. Additionally, the significant theory of mind deficit that patients with AD present (Moreau et al., 2016) makes it particularly challenging for them to attribute knowledge to their interlocutor in a real social interaction and detect perspectives that are different from their own. Thus, cases of incompatibility in terms of aspectual (grammatical and lexical) and temporal information can turn out to be very problematic. Therefore, a deeper understanding of the nature of the impairment in AD is very crucial given its essential role in communication and its special nature of bridging linguistic with extra-linguistic knowledge.

Related to this is the fact that languages mark tense and aspect in different ways. For instance, in English, grammatical aspect is encoded in auxiliaries and inflections (*I walked* vs. *I was walking*) while in Greek inflectional suffixes, infixes as well as stem changes denote a change in aspect (*e-graf-a* vs. *e-grap-s-a*). Cross-linguistic differences of this sort suggest that not only the conceptual knowledge, usage and perception of aspect might pose a problem for populations with cognitive decline but also its realization in specific languages. Taken all this together, the current contribution to "Words in the World" aims to highlight the multi-dimensional complexity of words, such as Greek verbs, in conveying subtle meanings which are encoded in standard linguistic tools, such as morphemes, and which might interact in various ways with the roots they attach to and with the sentential environment they appear in.

METHODOLOGY

Participants

Nine individuals with no neurological impairments (three males, six females, aged 70–85, MEAN: 79.5), 11 MCI (five males, six females, aged 65–84, MEAN: 73.8), and 11 mAD (five males, six females, aged 73–84, MEAN: 78), all native Greek-speakers, participated in the study. Participants were recruited from the Center of Physical Medicine and Rehabilitation in the area of Ioannina, Day Care Institution in the area of Ioannina and Larisa, and from the Laboratory of Logopathology of Technological Educational Institute of Epirus, in Greece. All were diagnosed by a qualified neurologist (GN).

The diagnosis of MCI was made in accordance with Petersen's criteria (Petersen et al., 2013). According to Petersen et al. (2013), the diagnostic criteria of MCI are: (1) memory complaints, (2) intact activities of daily living, (3) a score of 1.5 SD below the mean on neuropsychological measures (which is considered to be the standard cut-off point between healthy subjects and subjects with cognitive deficits), (4) Clinical Dementia Rating (CDR) from 0 to 0.5, (5) no dementia, (6) impairment in at least one cognitive domain (e.g., complex attention, executive function, learning, memory), and (7) general cognitive function (MoCa score from 20 to 25).

Regarding mAD, participants were included in the study if they fulfilled the following criteria: (1) a diagnosis of AD according to the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's and Related Disorders Association (NINCDS-ADRDA), (2) Clinical Dementia Rating (CDR) score = 1, (3) impairment in at least two cognitive domains (e.g., complex attention, executive function, learning, memory), and (4) general cognitive function (MoCA score from 14 to 20).

Neuropsychological examination revealed that patients did not suffer from any other (1) major psychiatric disorders (e.g., psychotic symptoms or disorders, alcohol or illegal drug abuse, depression), (2) neurological disorder (e.g., stroke, epilepsy, traumatic brain injury), and (3) visual/hearing impairment or writing/reading disability sufficient to impair performance in the assessment.

All participants had undergone clinical neurological assessment, blood tests and brain magnetic resonance imaging scans that presented no evidence of other diseases. To collect more information about the cognitive, functional and linguistic profile of the participants, we administrated additional neuropsychological tasks, translated and adapted versions for Greek. MoCA test (Nasreddine et al., 2005) was conducted to detect cognitive decline (working memory, repetition, audiovisual skills, etc.). To measure participants' attention, speed of processing and executive functions (task switching, ability to execute and modify a plan of action, ability to maintain two trails of thought simultaneously) the Trail Making tests (part A and part B) (Reitan and Wolfson, 1993) were used. Attention span and working memory capacity were evaluated with the Forward and Backward Digit Span (Sattler and Ryan, 2009; Holdnack, 2019). Language abilities were tested by a verbal fluency semantic task. The Boston Naming Test (Kaplan et al., 1983) was used to measure participants' naming abilities. Participants performed below the expected score in most of the tasks, suggesting the presence of cognitive dysfunction. Independent samples *t*-tests and Mann-Whitney test revealed that participants did not differ significantly in terms of mean age and education. Significant differences between the groups emerged in MoCA, Verbal Fluency Semantic task and Trail Making Test (part A) suggesting differences between MCI and mAD groups in cognitive decline, executive function and language abilities. In all other cognitive measurements no difference in their performance was observed. Participants' demographic and neuropsychological information, as well as statistical comparisons are presented in **Table 3**.

Experimental Tasks

A picture-naming task and a sentence-completion task were conducted. We scored correct responses, calculated percentages and also performed error analyses.

Naming Task Design

For the naming task 100 colored pictures were gathered from online sources. There were 20 pictures for each verb category (activity, accomplishment, state, semelfactive, and achievement). In order to avoid visual complexity that could affect participants' performance we only used pictures that depict one or maximum two people for all the five verb categories. Moreover, no background actions or objects, that could disorient participants,

TABLE 3 | Participants' demographic and neuropsychological mean scores (with standard deviations) and their statistical comparisons by using *t*-tests (for normally distributed data) and Mann-Whitney tests (for not normally-distributed data).

	MCI	mAD	Control group	MCI vs. mAD	mAD vs. Control group	MCI vs. Control group
Mean age	73.8	78	79.5	t = −1.832 $p = 0.082$ df = 20	t = −0.675 $p = 0.508$ df = 18	t = −2.06, $p = 0.054$ df = 18
Education	8.4 (3.0)	9.6 (3.7)	6.7 (1.5)	U = 50 $p = 0.458$ r = 0.16	U = 28 $p = 0.065$ r = 0.41	U = 35 $p = 0.195$ r = 0.33
			Statistical comparisons			
MoCA	22.0 (1.1)	17.0 (1.4)	U = 0, $p < 0.01$ r = 0.85			
Boston Naming	13.6 (1.2)	12.3 (1.8)	t = −1.905 $p = 0.071$ df = 20			
Verbal Fluency Semantic	32.9 (9.4)	23.7 (8.7)	t = −2.360 $p = 0.029$ df = 20			
TMTa	108.1 (36.6)	171.4 (55.4)	t =3.158 $p < 0.01$ df = 20			
TMTb	266.6 (52.4)	290.9 (30.1)	U = 86 $p = 0.065$ r = 0.40			
Backward digit span	5.9 (0.94)	5.9 (1.9)	U = 53.5 $p = 0.638$ r = 0.10			
Forward digit span	4.2 (1.0)	3.7 (1.1)	U = 38.5 $p = 0.133$ r = 0.32			

were illustrated. Also, four graduate students performed the task before the experimental groups of participants, in order to ensure that the pictures were recognized easily. We used Microsoft PowerPoint to present each picture separately to the participants.

Sentence-Completion Task

The sentence-completion task included 100 source sentences (SS) and target sentences (TS) pairs: 50 sentences were designed to test perfective aspect and 50 imperfective aspect. For each verb category (activity, state, achievement, semelfactive, and accomplishment) 20 pairs of sentences were constructed, 10 targeted the perfective aspect and 10 the imperfective. Only past and future tenses were used for the sentences with the majority of them being in past tense [79 past (38 imperfective, 41 perfective), 21 future (12 imperfective, nine perfective)]. Concerning time reference, past and future were distributed in the five verb categories as follows: activities (past 13, future seven), accomplishments (past 13, future seven), achievements (past 15, future five), states (past 19, future one), semelfactives (past 19, future 1). (2) and (3) are examples of perfective and imperfective aspect, respectively.

(2)SS: *Àvrio, i María θa potízi ton cípo óli méra.*
"Tomorrow, Maria will be watering $_{\text{3rd imperfective}}$ all day"
TS: *Àvrio, i María θa potísi ton cípo mésa se mía óra.*
"Tomorrow, Maria will water $_{\text{3rd perfective}}$ the garden within an hour"
(3)SS: *Xθes, to pedí ðjávase mésa se mía óra.*
"Yesterday, the child studied $_{\text{3rd perfective}}$ within an hour."
TS: *Xθes, to pedí ðjávaze epí mía óra.*
"Yesterday, the child was studying $_{\text{3rd imperfective}}$ for an hour."

Stimuli

The same 100 verbs were used in both tasks (see **Appendix**). Materials were split in two lists (List 1 and List 2) for counterbalancing purposes and also to ease and shorten the tasks in order to make them more suitable for brain-damaged participants. Each list contained 50 verbs (10 per lexical aspect verb category). Verbs were matched for number of syllables and number of letters by performing a *t*-test of independent samples ($p > 0.05$, in all comparisons). Frequency and argument structure were also taken into account and matched when possible. Concerning frequency, a *t*-test of independent samples revealed that state verbs were more frequent compared to semelfactives (t = −3.881, $p < 0.01$, d = 1.22), achievements (t = −2.355, $p = 0.024$, d = 0.74) and accomplishments (t = −4.544, $p < 0.01$, d = 1.43). Activities were more frequent than accomplishments (t = −2.117, $p = 0.041$, d = 0.68). All other verb categories were matched for frequency as well. In terms of argument structure, with the exception of state verbs and semelfactives, there is no difference between accomplishment, achievement and activity verbs. More specifically, all accomplishment (20) and all achievement (20) verbs were transitive, which means that they require at least two arguments (subject and object) to predicate their semantic and syntactic properties. Concerning activity verbs, 17 of the 20 verbs were transitive and the rest were intransitive. State verbs only included intransitive verbs, while semelfactives consisted of 11 intransitive verbs and 9 transitive verbs.

Imageability Ratings

Data for imageability and age of acquisition were collected by creating web-based questionnaires (Google forms). We collected data from 24 participants (eight males, 16 females), all native Greek-speakers, aged 18–35, University graduates, some with postgraduate degrees. The data were obtained following the instructions by Paivio et al. (1968), as they were presented in Rofes et al. (2018). Participants were instructed to rate a list of 100 words with respect to the ease or difficulty with which they arouse mental images based on their estimation. A 7-point scale was used, with one indicating low imageability rating, while seven indicating high imageability rating. Values of 2–6 indicated intermediate ratings. Statistical analysis (non-parametric Mann-Whitney test of two independent variables) revealed that activity verbs were rated significantly higher in terms of imageability, compared to all the other categories (activities vs. achievements: U = 69.5, $p < 0.01$, r = 0.55; activities vs. states: U = 83, $p < 0.01$, r = 0.50; activities vs. accomplishments: U = 85, $p < 0.01$, r = 0.49; activities vs. semelfactives: U = 108.5, $p < 0.01$, r = 0.39). Accomplishment verbs were rated significantly higher only compared to state verbs (U = 94, $p < 0.01$, r = 0.45). Achievements did not differ significantly in terms of imageability compared to states and semelfactives ($p > 0.05$ in all comparisons), with the latter being statistically higher compared to states (U = 118, $p = 0.026$, r = 0.35).

Age of Acquisition Ratings

We collected data from 28 participants (eight males, 20 females), all native Greek-speakers, aged 18–35, University graduates, some with postgraduate degrees. The data were obtained following the instructions by Dimitropoulou et al. (2009), as they were presented in Łuniewska et al. (2016). Participants were instructed to give an estimate of the age at which they thought they learned each of the 100 words in its written or oral form. A 5-point scale was used, with each number indicating the following age-bands: (1) 0–3 years, (2) 4–6 years, (3) 7–9 years, (4) 10–12 years and (5) at 13 years or later. Statistical analysis (non-parametric Mann-Whitney test of two independent variables) revealed no significant differences among verb categories ($p > 0.05$ in all comparisons), except for semelfactives which were acquired earlier compared to accomplishments (U = 108, $p < 0.01$, r = 0.25).

Procedure

The two lists were randomly assigned to participants and each participant was assigned one list only. Testing was completed in one session and it lasted approximately 20–25 min. The two tasks were presented to the participants in fixed order (picture-naming first followed by sentence completion) with a break in between.

Picture-Naming Task

Thirteen participants (5 MCI and 8 mAD) were examined on List 1 and 9 (6 MCI and 3 mAD) on List 2. PowerPoint was used to present each picture separately to the participants. Participants had to name the verb, which described the illustrated event by producing the 1st person singular of present tense. Instructions on how to complete the experimental task were provided at the beginning of the procedure. Four pictures that were not included in the stimuli were presented to participants in order to familiarize them with the task. Participants' responses during the trial period were not taken into account in the analysis. Participants had as much time as they needed in order to provide their answer. Each session lasted approximately 10–15 min.

Sentence Completion Task

Thirteen participants (5 MCI and 8 mAD) were examined on List 1 and 9 (6 MCI and 3 mAD) on List 2. Experimental materials were presented cross-modally to the participants who saw them on the computer screen and also heard the experimenter reading them aloud. Participants were asked to complete the missing verb from the TS in the correct form of grammatical aspect. At the beginning of the experimental procedure, participants were provided with instructions of how to complete the task. Four pairs of sentences that were not included in the stimuli were used as examples in order familiarize participants with the procedure. Participants' responses during the trial period were not taken into account in the analysis. The task was not chronometrized and participants had as much time as they needed in order to complete each sentence. Each session lasted approximately 10–12 min.

Scoring

For both picture-naming and sentence-completion task we performed quantitative and qualitative analyses taking into account participants' responses. This consisted of measuring percentages of correct responses as well as an error analysis in order to reveal error patterns.

In the sentence-completion task, for the quantitative analysis, we counted as correct those responses that contained the target verb in the correct aspectual form. When a mAD or an MCI participant completed a sentence using the verb in the target grammatical aspect (perfective or imperfective) but not in the correct person for the sentence to be grammatically acceptable, we considered that answer correct (e.g., Àvrio, i María θa potíz**un** ton cípo óli méra "*Tomorrow, Maria will be watering* $_{\textit{3rd plural imperfective}}$ *the garden all day*" instead of Àvrio, i María θa potíz**i** ton cípo óli méra "*Tomorrow, Maria will be watering* $_{\textit{3rd singular imperfective}}$ *the garden all day*"). Our goal was to examine participants' ability to produce the right type of grammatical aspect (perfective or imperfective) and not their ability to produce agreement. Similarly, different morphological forms of the same verb (e.g., χtipúsan "*They were hitting*" —χtípaγan "*They were hitting*") that encoded the target aspectual value were taken as correct answers. Responses that contained wrong use of tense (e.g., Ótan ímun mikrí, eγó zoγrafízo 1st singular present imperfective sinéχia "*When I was young, I am drawing all the time*" instead of Ótan ímun mikrí, eγó zoγráfiza 1st singular past imperfective sinéχia "*When I was young, I was drawing all the time*") were considered to be incorrect even when the aspectual characteristics of the verb were the targeted ones. For the error analysis, we checked for substitutions of aspect (e.g., perfective instead of imperfective and vice versa) as well as for time substitutions (e.g., past tense instead of future and vice versa).

In the picture-naming task, we excluded pictures that the control group did not recognize (six pictures from both List 1 and List 2 were excluded, out of which 4 depicted state verbs, one achievement verb and 1 semelfactive verb). We scored the ability of recalling and producing verbs and not the categories of agreement and tense, thus, responses that were in a different than the present tense (e.g., spáo "*I break*" —éspasa "*I broke*") were considered to be correct. Similarly, responses where participants used a different than the first person singular (e.g., spáme "*We break*" instead of spáo "*I break*" -) were not regarded as incorrect. Finally, we counted as correct responses that included a prefix (kliðóno "*I lock*" —ksekliðóno "*I unlock*"), provided that the produced word belonged to the same verb category and shared the same semantic and temporal features as the target, independently of change in meaning. With respect to error analysis, we checked for errors such as anomia (no response at all), responses unrelated to the target word, phonemic paraphasias and semantic paraphasias, which are common in naming tasks.

Statistics

In both tasks, logistic regression models were used to examine how *aspect* (grammatical vs. lexical), *tense* (past vs. future) and *group of participants* (control, MCI and AD) contributed to participants' responses (correct or wrong) and interacted with one another. When needed, additional non-parametric tasks were conducted to explore within group differences with respect to the choice of specific aspectual categories and tense.

We also performed correlation analyses by using the non-parametric Spearman rank test in order to measure the degree of association between participants' cognitive or general language abilities (e.g., results of MoCA, Boston Naming Test etc.) and their accuracy performance on the experimental tasks (sentence-completion task, picture-naming task). Given that both AD and MCI are primarily conditions that affect general cognition (see Introduction), participants' limited abilities might affect their accuracy in the experimental tasks. Thus, finding out whether there is an association between participants' performance and cognitive and language abilities can provide us with additional information and help interpret our results in a more comprehensive way. Finally, additional correlation analyses were performed by using the non-parametric Spearman rank test in order to measure possible associations between the lexical properties of imageability and age of acquisition and participants' accuracy scores in each verb category in the picture-naming task.

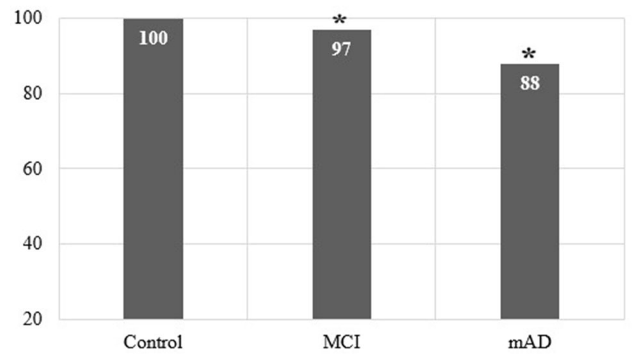


FIGURE 1 | Percentages of correct responses in grammatical aspect for the three groups of participants. The asterisk indicates significant effects ($p < 0.05$).

Results

No difference between control group performance in List 1 and List 2 was observed in both picture-naming and sentence-completion task (naming task: U = 21927.5, $p > 0.05$, r = 0, sentence completion task: U = 25,000, $p > 0.05$, r = 0). The same holds for the mAD group and their performance on both tasks (naming task: U = 26,148, $p = 0.794$, r = 0.01, sentence-completion task: U = 29,100, $p = 0.332$, r = 0.04). Concerning MCI, participants who were tested on List 1 performed significantly better on the picture-naming task compared to those who were tested on List 2 (U = 27,570, $p < 0.01$, r = 0.20), while in the sentence-completion task participants who were tested on List 2 performed better compared to those who tested on List 1 (U = 36,075, $p = 0.013$, r = 0.10). However, based on the fact that the other two groups did not show any difference on their performance and the fact that there was no clear dissociation between List 1 and List 2 among the MCI individuals' groups, we decided to move on an overall review of the results.

Sentence Completion Task

Overall results are shown in **Figure 1** while the percentages of correct responses by group, tense, lexical and grammatical aspect are shown in **Table 4**.

In all analyses, a binary logistic regression was performed treating participants' response (correct vs. incorrect) as depedent variable and *group*, *grammatical aspect*, *lexical aspect* and *tense* as predictors. The outcome is presented in **Table 5**. The results from the statistical model indicate that participants behave according to their group. That is, group is a significant predictor for their performance. The coefficient for Group (taking Controls as reference value) has a Wald statistic equal to 27.599 which is significant and the 0.001 level[df = 2]. When performing bootstrapping, both MCI and mAD are significant predictors too. The significance of B for both the MCI group (−19.196) and mAD (−17.690) is $p = 0.001$.

There are no other statistical predictors. That is, lexical aspect (*Wald* = 0.129, *p* = 0.720, *df* = 1), grammatical aspect (*Wald* =

0.040, $p = 0.841$, $df = 1$), and tense ($Wald = 0.001$, $p = 0.975$, $df = 1$), cannot predict participants' performance. Moreover, there is no statistically significant interaction between grammatical aspect and tense ($Wald = 0.006$, $p = 0.939$, $df = 1$), between lexical aspect and tense ($Wald = 0.010$, $p = 0.919$, $df = 1$), and between grammatical aspect and lexical aspect ($Wald = 0.838$, $p = 0.360$, $df = 1$).

The model explains 17% of the variability (Nagelkerke R square = 0.178) and it correctly predicted 100% of the correct answers and 0% of the incorrect answers, giving an overall percentage of correct prediction rate of 94.8%.

Finally, we found no correlation between patients' scores in the neuropsychological tasks and their performance in the sentence-completion. For individuals with mAD, analysis revealed the following values of no statistical significance: MoCA Rs (9) = −0.096, $p = 0.779$; Boston Naming Test Rs (9) = −0.353, $p = 0.286$; Verbal Fluency Semantic Task Rs (9) = −0.104, $p = 0.761$; Trail Making Test (part a) Rs (9) = 0.051, $p = 0.602$; Trail Making Test (part b) Rs (9) = 0.177, $p = 0.883$; Backward Digit Span test Rs (9) = −0.126, $p = 0.712$; Forward Digit Span test Rs (9) = −0.021, $p = 0.950$. Concerning the MCI groups' analysis the following values of significance were found: MoCA Rs (9) = 0.068, $p = 0.843$; Boston Naming Test Rs (9) = 0.089, $p = 0.794$; Verbal Fluency Semantic Task Rs (9) = 0.036, $p = 0.916$; Trail Making Test (part a) Rs (9) = 0.579, $p = 0.062$; Trail Making Test (part b) Rs (9) = −0.405, $p = 0.217$; Backward Digit Span test Rs (9) = −0.486, $p = 0.130$; Forward Digit Span test Rs (9) = −0.087; $p = 0.799$.

TABLE 4 | Percentages of correct responses by group, tense, grammatical aspect and lexical aspect in sentence completion task.

		AD	MCI
Tense	Past	87.7	97.5
	Future	89.3	95.8
Grammatical aspect	Perfective	88.5	97
	Imperfective	87.5	97.5
Lexical aspect	Activities	95.8	95.3
	Accomplishments	88	97
	Achievements	85.5	97
	Semelfactives	82.7	96
	States	90	99

Picture-Naming Task

Overall results are shown in **Figure 2**, while the percentages of correct responses by group and lexical aspect are shown in **Table 6**. As with the sentence completion task, a binary logistic regression analysis was performed by using SPSS. Participants' response (correct vs. incorrect) was treated as dependent variable while *group* (controls, mAD, MCI) and *lexical aspect* (activities, accomplishments, achievements, semefactives and states) were treated as predictors. The outcome is presented in **Table 7**.

The results from the statistical model indicate that participants behave according to their Group and according to the lexical aspect of the verb they have to name. The coefficient for Group variable (taking controls as reference value) has a Wald statistic

TABLE 5 | Outcome of logistic regression for sentence completion task.

		B	S.E.	Wald	df	Sig.	Exp(B)	95% CI for EXP(B)	
								Lower	Upper
Variables in the equation									
Step 1[a]	group			27.599	2	0.000			
	Group (1)	−19.196	1888.369	0.000	1	0.992	0.000	0.000	.
	Group (2)	−17.690	1888.369	0.000	1	0.993	0.000	0.000	.
	gram.asp by lex.asp	−0.156	0.170	0.838	1	0.360	0.856	0.613	1.194
	lex.asp by tense	0.029	0.288	0.010	1	0.919	1.029	0.586	1.809
	gram.asp by tense	−0.046	0.604	0.006	1	0.939	0.955	0.292	3.120
	lex.asp	1.175	3.278	0.129	1	0.720	3.239	0.005	1998.537
	gram.asp	1.241	6.187	0.040	1	0.841	3.460	0.000	639449.570
	Tense	0.172	5.499	0.001	1	0.975	1.187	0.000	56919.279
	Constant	12.198	1889.203	0.000	1	0.995	198392.877		

	Bootstrap					
	B	Bias	Std. error	Sig. (2–tailed)	95% confidence interval	
					Lower	Upper
Group (1)	−19.196	0.003	0.137	0.001	−19.446	−18.910
Group (2)	−17.690	0.043	0.264	0.001	−18.106	−17.055

[a]*Variable(s) entered on step 1: group, gram.asp * lex.asp, lex.asp * tense, gram.asp * tense, lex.asp, gram.asp, tense. Variable coding is translated as follows: Group = control group (reference value), Group (1) = MCI, (Group 2) = mAD.*

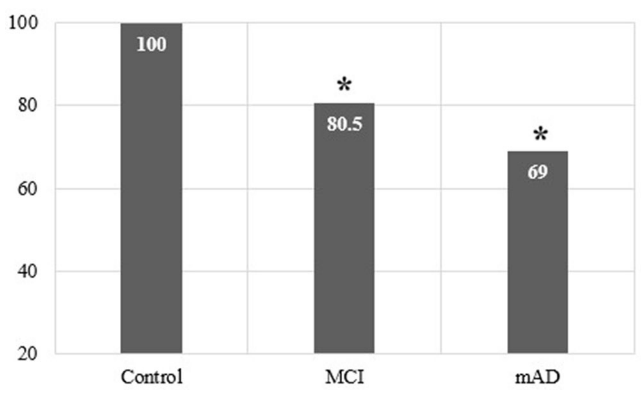


FIGURE 2 | Percentages of correct responses in picture-naming task for control group, individuals with MCI and individuals with mAD. The asterisk indicates significant effects ($p < 0.05$).

TABLE 6 | Percentages of correct responses by group and lexical aspect in the naming task.

	AD	MCI
Activities	82	91
Accomplishments	79	86
Achievements	49	74
Semelfactives	70.5	76
States	63	73

equal to 15.807 which is significant and the 0.001 level[df = 2]. When performing bootstrapping, both MCI and mAD are significant predictors too. The significance of B for both the MCI group (−19.791) and mAD (−20.392) is $p = 0.001$.

Similarly, *lexical aspect* is also a statistical predictor. When taking *achievements* as reference value, *activities* and *accomplishments* yield significantly better naming (Wald = 31.351, $p < 0.001$ and Wald = 25.502, $p < 0.001$, respectively), *semelfactives* yield marginally better naming (Wald = 4.026, $p = 0.045$), while state verbs clearly yield worse naming than the previous categories but not compared to achievements (Wald = 1.358, $p = 0.244$). No significant interaction between Group and Lexical aspect has emerged.

The model explains 27.3% of the variability (Nagelkerke R square = 0.273) and it correctly predicted 100% of the correct answers and 0% of the incorrect answers, giving an overall percentage of correct prediction rate of 82.6%.

No significant correlations between patients' scores on neuropsychological tasks and their performance in the naming task were observed. Correlation analyses for individuals with mAD revealed the following values of no statistical significance: MoCA Rs (9) = −0.216; $p = 0.524$, Boston Naming Test Rs (9) = −0.135, $p = 0.693$; Verbal Fluency Semantic Task Rs (9) = 0.344, $p = 0.300$; Trail Making Test (part A) Rs (9) = 0.389, $p = 0.238$; Trail Making Test (part b) Rs (9) = −0.202, $p = 0.551$; Backward Digit Span test Rs (9) = −0.245, $p = 0.467$; Forward Digit Span test Rs (9) = −0.433, $p = 0.184$. Regarding individuals with MCI the following values of significance were found: MoCA Rs (9) = −0.185, $p = 0.584$; Boston Naming Test Rs (9) = −0.434, $p = 0.183$; Verbal Fluency Semantic Task Rs (9) = −0.327, $p = 0.326$; Trail Making Test (part A) Rs (9) = −0.028, $p = 0.936$; Trail Making Test (part b) Rs (9) = −0.074, $p = 0.818$, Backward Digit Span test Rs (9) = 0.333, $p = 0.317$; Forward Digit Span test Rs (9) = −0.028, $p = 0.935$. Moreover, we observed no correlation between the factors of frequency, imageability, age of acquisition and participants' performance in the picture-naming task. Specifically, for individuals with mAD a very weak negative correlation between frequency and participants' performance was observed, but it did not reach significance [Rs (92) = −0.127, $p = 0.222$]. Similarly, an insignificant very weak positive correlation between imageabilty and participant's score was found [Rs (92) = 0.156, $p = 0.132$]. Regarding the age of acquisition, a non-significant very weak negative correlation was observed [Rs (92) = −0.003, $p = 0.979$]. When it comes to individuals with MCI, analysis revealed the following positive or negative correlations [frequency: very weak positive correlation with no statistical significance Rs (92) = 0.126, $p = 0.228$; imageability: a weak positive correlation, which reached significance Rs (92) = 0.313, $p = 0.002$; age of acquisition: a very weak correlation, which did not reach significance Rs(92) = 0.168, $p = 0.106$].

Error analysis revealed that the most common mistake of both MCI and mAD participants (67.3 and 58.9%, respectively) were responses unrelated to the target verb (e.g., instead of káθome "*sit*" —pézo "*play*"), followed by semantic paraphasias (15.8%, and 17.3% of the cases, respectively). In the majority of semantic paraphasias, the target verb was substituted by another verb which either belonged to the same semantic category with the target (e.g., sfugarízo "*mop*" instead of skupízo– "*wipe*") or it was connected with it by hyponymy. In the hyponymy relationship there is a hypernym word, which constitutes the general category that includes the hyponym words. Usually participants used the hypernym word (e.g., tróo *"eat"* instead of ðagnóo "*bite*") when the target was a hyponym one. In few cases, participants used circumlocutions to name the target verb (e.g., káno bánjo *"take a swim"* instead of kolibáo "*swim*") (7.9% MCI and 8.9% mAD). Cases where participants did not respond at all were also observed in both groups (5.9% MCI and 10.2% mAD). Finally, incorrect responses also include cases in which the target word and the answer may not necessarily belong to the same category but relate to each other based on a scenario (e.g., nistázo "*I am sleepy*" instead of χazmurjéme "*I yawn*").

DISCUSSION

The aim of the current investigation was to contribute to the existing literature on *aspect* and time reference in neurodegenerative conditions by providing an account that takes into consideration temporal features such as *duration,* as a dual variable, with grammatical as well as lexical instantiations. In other words, we investigated how the temporal feature of *duration* either as functional-grammatical variable or as lexical variable affects participants' performance and whether

TABLE 7 | Outcome of logistic regression for the naming task.

		B	S.E.	Wald	df	Sig.	Exp(B)	95% CI for EXP(B)	
								Lower	Upper
Variables in the equation									
Step 1[a]	Group			15.807	2	0.000			
	Group (1)	−19.791	1921.475	0.000	1	0.992	0.000	0.000	.
	Group (2)	−20.392	1921.475	0.000	1	0.992	0.000	0.000	.
	Lexasp			47.300	4	0.000			
	Lexasp (1)	0.428	0.213	4.026	1	0.045	1.534	1.010	2.331
	Lexasp (2)	0.257	0.221	1.358	1	0.244	1.293	0.839	1.992
	Lexasp (3)	1.387	0.248	31.351	1	0.000	4.004	2.464	6.506
	Lexasp (4)	1.200	0.238	25.502	1	0.000	3.319	2.083	5.286
	Constant	20.623	1921.475	0.000	1	0.991	904211111.5		

	B	Bootstrap				
		Bias	Std. error	Sig. (2–tailed)	95% confidence interval	
					Lower	Upper
Group (1)	−19.791	0.004	0.116	0.001	−20.001	−19.534
Group (2)	−20.392	−0.003	0.107	0.001	−20.590	−20.170

[a]*Variable(es) entered on step 1: group, lexasp. Variable coding is translated as follows: Group = control group (reference value), Group (1) = MCI, (Group 2) = mAD; LexAsp = achievement (reference value), LexAsp (1) = semelfactive, LexAsp (2) =state, LexAsp (3) =activity, LexAsp (4) = accomplishment.*

the two interact. While clinical populations' ability to produce grammatical aspect either under the umbrella of functional features or as an indicator of time reference has been investigated extensively, their ability to produce lexical aspect has been widely neglected. Our goal was to provide a unitary account of aspect, if possible, and address the inconsistency of previous results with respect to grammatical aspect by gaining insights from lexical aspect. To this end, we used a picture-naming and a sentence-completion task to investigate the performance of Greek-speaking mAD and MCI individuals on using and naming verbs that differ in terms of their *lexical* and *grammatical aspect.*

As far as *grammatical aspect* is concerned, previous results have been controversial with no attempt to explain this inconsistency. Our data from the sentence-completion task suggest that *grammatical aspect* is impaired in individuals with mAD and MCI and are in line with other studies with mAD participants, i.e., Altmann et al. (2001), Fyndanis et al. (2013), and Roumpea et al. (2019). With respect to MCI participants, results are in line with Roumpea et al. (2019), who reported impaired grammatical aspect in this population but at odds with de Jager et al. (2003), who found MCI participants' syntactic abilities equally preserved as in control individuals.

A second important piece of evidence is the lack of difference between participants' preference for perfective vs. imperfective aspect. In other words, there was no significant preponderance of perfective (*I broke*) over imperfective (*I was breaking*). This is at odds with Fyndanis et al. (2013) and also with the assumption that unmarked features (imperfective) are better preserved than the marked features (perfective) (Lapointe, 1985). In contrast, the finding suggests that grammatical aspect is generally impaired in mAD independently of markedness or of ±duration. Thus, the prediction that duration, as functional-grammatical feature, would affect participants' performance leading them to prefer the perfective aspect even in imperfective contexts was not supported by the data.

When examining our results with respect to time reference, no interaction between time and grammatical aspect was found. That is, no preference of perfective over imperfective was found in past reference context and similarly no better performance was observed on imperfective in future reference context in either MCI or mAD groups. Participants performed equally well in perfective and imperfective aspect independently of the time reference context, ruling out the possibility that time reference is a factor that might interfere with participants' ability to produce grammatical aspect and choose between perfective or imperfective aspect (contra Dragoy and Bastiaanse, 2013 and in line with Fyndanis and Themistocleous, 2019).

Finally, no interaction between grammatical and lexical aspect emerged, with patients' performance not being affected by verb category when producing grammatical aspect. This lack of interaction suggests that the lexical aspect of the verb does not influence participants' choice of grammatical aspect (perfective vs. imperfective). If indeed lexical aspect was a decisive factor in choosing grammatical aspect, then mismatches in term of duration (a property of both lexical and grammatical aspect) would have intereferred with participants' performance. In other words, patients would have difficulties in producing the imperfective aspect (duration) both in inherently durative verbs (e.g., activities "*run*") and in verbs with instantaneous (no duration) meaning (e.g., semelfactives "*hit*"), a performance that was not observed in our data. The lack of interaction between lexical and grammatical aspect in sentence completion task is indicative of the independence of the two variables which can be affected differentially in populations with semantic and cognitive decline.

Given that no significant correlations were detected with the various neuropsychological variables (see Results), we assume that degraded working memory, verbal fluency and executive dysfunction of patients of this study are not responsible for their inability to correctly produce grammatical aspect. Of course, these factors cannot entirely be excluded, especially given that the small number of participants might have prevented us from detecting potentially significant correlations. However, we cannot base our interpretation on a hypothetical possible result. Thus, we are left with the possibility that the underlying reason for participants' failure lies either in the conceptual sphere of grammatical aspect or in a difficulty to materialize the concept of grammatical aspect by constructing the corresponding morphological forms. We will explore these two possibilities in turn.

Aspect has always been considered a demanding category. It is impaired in aphasic populations as well and it is a feature which is acquired at a later stage of language acquisition as it coincides with various other cognitive developmental factors (Clark, 2009). Thus, the connection between the system of aspect and higher cognitive functions has been pointed out multiple times. The main difficulty associated with aspect is its dual status as it marks the temporal contour of events by means of inherent lexical meanings (durative vs. non-durative verbs) but also with grammatical morphology. It is exactly in the cross-road of this double status, that problems start to emerge. In the sentence completion task, we found no interaction between lexical and grammatical aspect and also no difference between perfective and imperfective which suggests that grammatical aspect *on its own and in its entirety* poses difficulties for participants. This is indicative of an event conceptualization problem. Namely, participants are not able to detect how the specific events are realized in time, that is, either as events with internal perspective that highlight the gradual development (in case of imperfective) or as events with external perspective that focus on the end state and provide a "glimpse" of the whole event (perfective). Crucial to that is also the comprehension of adverbials that were used in target sentences as well as the theory of mind deficit which is common in AD (Moreau et al., 2016) and which prevents AD individuals from identifying with the perspective of their interlocutor and in our case, the perspective surrounding the source sentences.

The second possibility follows the logic that participants can grasp the internal temporal consistency of events but they have a problem with the implementation of their choice, that is, with the creation of the correct grammatical form that describes their choice. When it comes to aspect realization in Greek, several morphophonological operations have to be performed. As mentioned in section Grammatical aspect, in order to form the perfective aspect, one needs to add the aspectual marker -s- to the verbal stem, insert the augmentative vowel *e-* and also add an inflectional suffix such as *-a* for the first person singular (e.g., líno "*I solve*" → éli-**s**-a "*I solved*"). Thus, it is not impossible for individuals with mAD and MCI to have difficulties in reconstructing the morphological features connected with the expression of aspect. Even though previous studies have also discussed this possibility in light of impaired performance of tense and aspect (Bastiaanse et al., 2011; Fyndanis et al., 2013), this is an issue that calls for further investigation, especially given that mAD and MCI are not known to have serious problems with morphology itself.

Let us now discuss the results of *picture-naming task* which targeted the investigation of *lexical aspect*. MCI and mAD groups performed significantly lower than controls, with mAD being worse than MCI, suggesting that lexical aspect is impaired in both populations. These findings are in line with previous studies that report impaired recall abilities of verbs (Alegret et al., 2018) and naming difficulties (Druks et al., 2006; Masterson et al., 2007) in MCI and mAD. Despite the difference between the performance of MCI and mAD individuals, for both groups activities (*walk*) and accomplishments (*build*) were found to be better preserved compared to states (*know*), achievements (*break*) and semelfactive (*hit*) verbs. Thus, it appears that duration alone, as a lexical variable and as part of lexical representation of a verb is not a decisive factor when it comes to participants' choice. If indeed duration affected participants' performance, then either all durative verbs (states, activities, accomplishments) would have been impaired or all of them would have been better preserved.

Furthermore, correlation analyses have shown that the various psycholinguistic variables related to the stimulus set did not affect the final results with just one exception for MCI participants and high imageability of accomplishments and achievements. Specifically, the high frequency of state verbs did not lead to better performance on this category which was found to be one of the most impaired. Similarly, participants performed equally high in activities and accomplishments, even though activities were more frequent. Results were also free from any age of acquisition effects as the lack of significant correlations between participant's performance and this variable suggests. Finally, imageability did not affect the performance of mAD but there was a weak correlation between imageability and MCI participant's performance indicating that the high imageability ratings of accomplishment and activity verbs might have interfered with the results.

The fact that only *accomplishments* (*build*) and *activities* (*run*) were better preserved leads to the assumption that the key element of impaired performance in naming verbs might have to do not only with their internal duration but with the combination of duration with internal semantic complexity. Activities and accomplishments are the two types of verbs in this study which have internal structure (see section Lexical Aspect above). They both present processes that not only last in time but they also consist of different successive phases in which the processes evolve. This contrasts with *achievements* (*break*) which are instantaneous events with no duration and internal structure and also with states (*sleep*), which lack internal structure as they describe situations consisting of identical stages, even though they are durative. Thus, we assume that it is the combination of the semantic and temporal features of activities and accomplishments that make them more prominent and better preserved in populations with semantic limitations, such as MCI and mAD. In general, semantic complexity has been found to affect individuals' abilities to produce verbs both in aphasia and AD (Breedin et al., 1998; Kim and Thompson, 2004). In Breedin et al. (1998) aphasic patients were found better at retrieving semantically complex verbs compared to simpler ones (e.g., heavy verbs *"run"* vs. light verbs "*go,*". specific verbs "*wipe*" vs. general verbs "*clean*"). The authors argued that semantically complex verbs (e.g., *run*, *wipe*) contain rich semantic features that make their meaning more specific and thus they are more distinctive and easier to recall compared to semantically simpler verbs. This explains very nicely the pattern we have found in the current investigation.

To sum up, apart from contributing data about grammatical and lexical aspect in neurodegenerative diseases, the current study also aimed to look at aspect at the big picture possibly as a unified category (lexical + grammatical) associated with the temporal dimension of events. Our findings do not provide grounds for a unified account of aspect. In contrast, they strongly suggest the independence of the two subsystems, lexical and grammatical. In the heart of this argument is the temporal feature of duration which has a different effect depending on whether it is processed as a lexical feature or as a functional feature. That is, as a lexical feature, combined with internal structure, duration has a positive effect in increasing the verb's saliency, thus, making it easier for retrieval. As a functional feature encoded in grammatical aspect, duration does not seem to play any decisive role in participants' performance, leaving them in the dark when it comes to the choice between perfective and imperfective. This leaves open the possibility that the observed performance of mAD and MCI participants could be related to difficulties with formulating the morphological forms that encode perfective and imperfective aspect, something which is not necessary for the production of lexical aspect. Or it is indicative of a general inability to comprehend and integrate aspectual information in a sentence. Whatever the underlying reason might be, what we learned from the current study is that grammatical aspect in its essence can be impaired in neurodegeneration and this is independent of any other factors. Any claims, of course, should take into account the limitations of the current study, such as the variables that were not possible to control for that might have affected participants' performance.

Last but not least, we would like to emphasize the importance of studies that provide linguistic evidence from populations with cognitive decline, like MCI and mAD. An accurate description of patients' linguistic performance is a pivotal first step in securing appropriate intervention processes. As our knowledge from neurolinguistic research advances, data-driven intervention programs are becoming a necessity and nowadays are also becoming a reality. Similarly, diagnostic tools have slowly started incorporating evidence from linguistically informed studies. There is still a long way to go, but it is the only way we could possibly secure precise diagnosis and appropriate intervention. This is particularly important for populations who speak under-studied languages. In these cases, the majority of diagnostic and intervention tools is mostly adaptations or even direct translations from tools created for English-speaking populations. As a result, crucial features of any specific language which differ from English are often not taken into account. The current study, with its limitations notwithstanding, falls into this scope of providing additional evidence about subtle linguistic features for Greek-speaking populations which are often neglected by both diagnostic and intervening tools. As such, it offers a ground for its use for clinical purposes as well.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The research protocol was approved by the Ethics Committee of the Medical School of Larissa, University of Thessaly, and it was conducted in accordance with the principles of the Declaration of Helsinki. Written consent was obtained from all the participants (or their caregivers) after having been informed of the nature of the study they would take part in. However, when it comes to the participants from the Center of Physical Medicine and Rehabilitation in the area of Ioannina, the experimental tasks were conducted as part of their other daily activities and after being informed about the nature of the study they gave their oral consent.

AUTHOR CONTRIBUTIONS

CM conceived the project, designed the experiments, supervised the whole process from data collection, and analysis to interpretation. She also wrote the final version of the manuscript. GR was responsible for reviewing the literature on theoretical background and background research, conducting the experimental tasks to healthy individuals, analyzing the data, and writing the first draft of the manuscript. CM and GR were responsible for all revisions after the first round of reviewing. AN was responsible for collecting the data and conducting the experimental tasks to MCI and AD participants. SS provided feedback and comments concerning the analysis and the writing of the manuscript. GN identified the participants, performed the clinical diagnosis, and classified them to MCI and AD groups. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We wish to thank the participants for taking part in this study and the Institutions (Central of Physical Medicine and Rehabilitation in the area of Ioannina, Day Care Institution in the area of Ioannina and Larisa and the Laboratory of Speech

Therapy of Technological Educational Institute of Epirus) for giving us the permission to be involved with their treated patients and perform our experimental tasks. Parts of this research were presented at the 19th Science of Aphasia XIX Conference (2018) and 13th Alps Adria Psychology Conference (2018). We thank the audiences for their feedback. Finally, we would like to express our gratitude to Valantis Fyndanis and to Gary Libben for providing valuable feedback to earlier versions of the paper and to Artur Stepanov for help with the statistical analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2020.434106/full#supplementary-material

REFERENCES

Albert, M., Blacker, D., Moss, M. B., Tanzi, R., and McArdle, J. J. (2007). Longitudinal change in cognitive performance among individuals with mild cognitive impairment. *Neuropsychology* 21, 158–169. doi: 10.1037/0894-4105.21.2.158

Alegret, M., Peretó, M., Pérez, A., Valero, S., Espinosa, A., Ortega, G., et al. (2018). The role of verb fluency in the detection of early cognitive impairment in alzheimer's disease. *J. Alzheimer's Dis.* 62, 611–619. doi: 10.3233/JAD-170826

Alexopoulos, P., Grimmer, T., Perneczky, R., Domes, G., and Kurz, A. (2006). Progression to dementia in clinical subtypes of mild cognitive impairment. *Dement. Geriatr. Cogn. Disord.* 22, 27–34. doi: 10.1159/000093101

Altmann, L. G. P., Kempler, D., and Andersen, E. S. (2001). Speech errors in alzheimer's disease: reevaluating morphosyntactic preservation. *J. Speech Lang. Hear. Res.* 44, 1069–1082. doi: 10.1044/1092-4388(2001/085)

Appell, J., Kertesz, A., and Fisman, M. (1982). A study of language functioning in alzheimer patients. *Brain Lang.* 17, 73–91. doi: 10.1016/0093-934X(82)90006-2

Bastiaanse, R., Bamyaci, E., Hsu, C. J., Lee, J., Duman, T. Y., and Thompson, C. K. (2011). Time reference in agrammatic aphasia: a cross-linguistic study. *J. Neuroling.* 24, 652–673. doi: 10.1016/j.jneuroling.2011.07.001

Bayles, K. A. (1982). Language function in senile dementia. *Brain Lang.* 16, 265–280. doi: 10.1016/0093-934X(82)90086-4

Braaten, A. J., Parsons, T. D., McCue, R., Sellers, A., and Burns, W. J. (2006). Neurocognitive differential diagnosis of dementing diseases: alzheimer's dementia, vascular dementia, frontotemporal dementia, and major depressive disorder. *Int. J. Neurosci.* 116, 1271–1293. doi: 10.1080/00207450600920928

Breedin, S. D., Saffran, E. M., and Schwartz, M. F. (1998). Semantic factors in verb retrieval: an effect of complexity. *Brain Lang.* 63, 1–31. doi: 10.1006/brln.1997.1923

Chomsky, N. (1995). *The Minimalist Program (Vol. 28)*. Cambridge, MA: MIT press.

Chomsky, N. (2000). "Minimalist inquiries: the framework," in *Step by Step,* eds. R. Martin, D. Michaels, and J. Uriagereka (Cambridge, MA: MIT Press), 89–155.

Chomsky, N. (2001). "Derivation by phase," in *Ken Hale: A Life in Language,* ed. K. Kenstowicz (Cambridge, MA: MIT Press), 1–52.

Clark, E. (2009). First *Language Acquisition*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511806698

Comrie, B. (1976). *Aspect*. New York, NY: Cambridge University Press.

de Jager, C. A., Hogervorst, E., Combrinck, M., and Budge, M. M. (2003). Sensitivity and specificity of neuropsychological tests for mild cognitive impairment, vascular cognitive impairment and alzheimer's disease. *Psychol. Med.* 33, 1039–1050. doi: 10.1017/S0033291703008031

Demetriou, E., and Holtzer, R. (2017). Mild cognitive impairments moderate the effect of time on verbal fluency performance. *J. Int. Neuropsychol. Soc.* 23, 44–55. doi: 10.1017/S1355617716000825

Dimitropoulou, M., Duñabeitia, J. A., Blitsas, P., and Carreiras, M. (2009). A standardized set of 260 pictures for Modern Greek: norms for name agreement, age of acquisition, and visual complexity. *Behav. Res. Methods* 41, 584–589. doi: 10.3758/BRM.41.2.584

Dragoy, O., and Bastiaanse, R. (2013). Aspects of time: time reference and aspect production in Russian aphasic speakers. *J. Neurolinguist.* 26, 113–128. doi: 10.1016/j.jneuroling.2012.05.003

Druks, J., Masterson, J., Kopelman, M., Clare, L., Rose, A., and Rai, G. (2006). Is action naming better preserved (than object naming) in alzheimer's disease and why should we ask? *Brain Lang.* 98, 332–340. doi: 10.1016/j.bandl.2006.06.003

Duong, A., Whitehead, V., Hanratty, K., and Chertkow, H. (2006). The nature of lexico-semantic processing deficits in mild cognitive impairment. *Neuropsychologia* 44, 1928–1935. doi: 10.1016/j.neuropsychologia.2006.01.034

Fyndanis, V., Arfani, D., Varlokosta, S., Burgio, F., Maculan, A., Miceli, G., et al. (2018). Morphosyntactic production in Greek-and Italian-speaking individuals with probable alzheimer's disease: evidence from subject–verb agreement, tense/time reference, and mood. *Aphasiology* 32, 61–87. doi: 10.1080/02687038.2017.1358352

Fyndanis, V., Manouilidou, C., Koufou, E., Karampekios, S., and Tsapakis, E. M. (2013). Agrammatic patterns in alzheimer's disease: evidence from tense, agreement, and aspect. *Aphasiology* 27, 178–200. doi: 10.1080/02687038.2012.705814

Fyndanis, V., and Themistocleous, C. (2019). Are there prototypical associations between time frames and aspectual values? Evidence from Greek aphasia and healthy aging. *Clin. Linguist. Phonet.* 33, 191–217. doi: 10.1080/02699206.2018.1480657

Holdnack, J. A. (2019). "The development, expansion, and future of the WAIS-IV as a cornerstone in comprehensive cognitive assessments," in *Handbook of Psychological Assessment*, eds G. Goldstein, D. N. Allen, and J. DeLuca (Hoboken, NJ: Academic Press), 103–139.

Holton, D., Mackridge, P.και Φιλιππακη-, and Warburton. (2010). *Grammar of Greek Language (14th Edition). Spiropoulos Vasilis (Translation in Greek)*. Athens: Patakis Publications.

Kaplan, E. F., Goodglass, H., and Weintraub, S. (1983). *Boston Naming Test*. Philadelphia, PA: Lea & Febiger.

Kaprinis, S., and Stavrakaki, S. (2007). Morphological and syntactic abilities in patients with alzheimer's disease. *Brain Lang.* 103, 59–60. doi: 10.1016/j.bandl.2007.07.044

Kavé, G., and Dassa, A. (2018). Severity of alzheimer's disease and language features in picture descriptions. *Aphasiology* 32, 27–40. doi: 10.1080/02687038.2017.1303441

Kavé, G., and Goral, M. (2018). Word retrieval in connected speech in alzheimer's disease: a review with meta-analyses. *Aphasiology* 32, 4–26. doi: 10.1080/02687038.2017.1338663

Kavé, G., and Levy, Y. (2003). Morphology in picture descriptions provided by persons with alzheimer's disease. *J. Speech Lang. Hear. Res.* 46, 341–352. doi: 10.1044/1092-4388(2003/027)

Kim, M., and Thompson, C. K. (2004). Verb deficits in alzheimer's disease and agrammatism: Implications for lexical organization. *Brain Lang.* 88, 1–20. doi: 10.1016/S0093-934X(03)00147-0

Kok, P., van Doorn, A., and Kolk, H. (2007). Inflection and computational load in agrammatic speech. *Brain Lang.* 102, 273–283. doi: 10.1016/j.bandl.2007.03.001

Koukoulioti, V. (2013). *Argument structure and verb inflection: a comparison between Greek-speaking patients with aphasia and semantic dementia (Doctoral* dissertation), Thessaloniki: Aristotle University of Thessaloniki.

Lambon Ralph, M. A., Patterson, K., Graham, N., Dawson, K., and Hodges, J. R. (2003). Homogeneity and heterogeneity in mild cognitive impairment and alzheimer's disease: a cross-sectional and longitudinal study of 55 cases. *Brain* 126, 2350–2362. doi: 10.1093/brain/awg236

Lapointe, S. G. (1985). A theory of verb form use in the speech of agrammatic aphasics. *Brain Lang.* 24, 100–155. doi: 10.1016/0093-934X(85)90100-2

Lindeboom, J., and Weinstein, H. (2004). Neuropsychology of cognitive ageing, minimal cognitive impairment, Alzheimer's disease, and vascular cognitive impairment. *Euro. J. Pharmacol.* 490, 83–86. doi: 10.1016/j.ejphar.2004.02.046

Łuniewska, M., Haman, E., Armon-Lotem, S., Etenkowski, B., Southwood, F., Andelkovi,ć, D., et al. (2016). Ratings of age of acquisition of 299 words across 25 languages: is there a cross-linguistic order of words? *Behav. Res. Methods* 48, 1154–1177. doi: 10.3758/s13428-015-0636-6

Manouilidou, C., and de Almeida, R. G. (2009). "Linguistic canonicity and verb deficits in alzheimer's disease," in *The Fruits of Empirical Linguistics I: Process,* eds. S. Featherston and S. Winkler (Berlin: De Gruyter), 123–150. doi: 10.1515/9783110216141.123

Manouilidou, C., de Almeida, R. G., Schwartz, G., and Nair, N. P. V. (2009). Thematic roles in alzheimer's disease: hierarchy violations in psychological predicates. *J. Neurolinguist.* 22, 167–186. doi: 10.1016/j.jneuroling.2008.10.002

Manouilidou, C., Dolenc, B., Marvin, T., and Pirtošek, Z. (2016). Processing complex pseudo-words in mild cognitive impairment: the interaction of preserved morphological rule knowledge with compromised cognitive ability. *Clin. Linguist. Phonet.* 30, 49–67. doi: 10.3109/02699206.2015.1102970

Masterson, J., Druks, J., Kopelman, M., Clare, L., Garley, C., and Hayes, M. (2007). Selective naming (and comprehension) deficits in alzheimer's disease? *Cortex* 43, 921–934. doi: 10.1016/S0010-9452(08)70691-9

Moreau, N., Rauzy, S., Viallet, F., and Champagne-Lavau, M. (2016). Theory of mind in alzheimer disease: evidence of authentic impairment during social interaction. *Neuropsychology* 30, 312–321. doi: 10.1037/neu0000220

Moser, A. (2013). "Aspect and aktionsart: a study on the nature of grammatical categories," in Major Trends in Theoretical and Applied Linguistics: Selected Papers from the 20th International Symposium on Theoretical and Applied Linguistics (April 1-3, 2011), eds. N. Lavidas, T. H. Alexiou and A. Sougari (Berlin: Versita de Gruyter), 99–121.

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., et al. (2005). The montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* 53, 695–699. doi: 10.1111/j.1532-5415.2005.53221.x

Paivio, A., Yuille, J. C., and Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *J. Exp. Psychol.* 76, 1–25. doi: 10.1037/h0025327

Petersen, R. C. (2003). *Mild Cognitive Impairment: Aging to Alzheimer's Disease*. New York, NY: Oxford University Press.

Petersen, R. C., Aisen, P., Boeve, B. F., Geda, Y. E., Ivnik, R. J., Knopman, D. S., et al. (2013). Mild cognitive impairment due to Alzheimer disease in the community. *Annals Neurol*. 74, 199–208. doi: 10.1002/ana.23931

Petersen, R. C., Doody, R., Kurz, A., Mohs, R. C., Morris, J. C., Rabins, P. V., et al. (2001). Current concepts in mild cognitive impairment. *Arch. Neurol.* 58, 1985–1992. doi: 10.1001/archneur.58.12.1985

Reitan, R. M., and Wolfson, D. (1993). *The Halstead-Reitan Neuropsychological Test Battery: Theory and Clinical Interpretation*, 2nd Edn. Tucson, AZ: Neuropsychology Press.

Rofes, A., Zakariás, L., Ceder, K., Lind, M., Johansson, M. B., De Aguiar, V., et al. (2018). Imageability ratings across languages. *Behav. Res. Methods* 50, 1187–1197. doi: 10.3758/s13428-017-0936-0

Roumpea, G., Nousia, A., Stavrakaki, S., Nasios, G., and Manouilidou, C. (2019). Lexical and grammatical aspect in mild cognitive impairment and alzheimer's disease. *Select. Papers Theoretical Appl. Linguist.* 23, 381–397. doi: 10.26262/istal.v23i0.7355

Salehi, M., Reisi, M., and Ghasisin, L. (2017). Lexical retrieval or semantic knowledge which one causes naming errors in patients with mild and moderate alzheimer's disease. *Dement. Geriatr. Cogn. Disord. Extra* 7, 419–429. doi: 10.1159/000484137

Sattler, J. M., and Ryan, J. J. (2009). *Assessment with the WAIS-IV*. San Diego, CA: Jerome M Sattler Publisher.

Silagi, M. L., Bertolucci, P. H. F., and Ortiz, K. Z. (2015). Naming ability in patients with mild to moderate alzheimer's disease: what changes occur with the evolution of the disease? *Clinics* 70, 423–428. doi: 10.6061/clinics/2015(06)07

Smith, C. (1997). *The Parameter of Aspect (2nd ed)*. Dordrecht: Kluwer Academic Publishers. doi: 10.1007/978-94-011-5606-6

Summers, M. J., and Saunders, N. L. (2012). Neuropsychological measures predict decline to alzheimer's dementia from mild cognitive impairment. *Neuropsychology* 26, 498–508. doi: 10.1037/a0028576

Taler, V., and Phillips, N. A. (2008). Language performance in alzheimer's disease and mild cognitive impairment: a comparative review. *J. Clin. Exp. Neuropsychol.* 30, 501–556. doi: 10.1080/13803390701550128

Visser, P. J., Scheltens, P., Verhey, F. R., Schmand, B., Launer, L. J., Jolles, J., et al. (1999). Medial temporal lobe atrophy and memory dysfunction as predictors for dementia in subjects with mild cognitive impairment. *J. Neurol.* 246, 477–485. doi: 10.1007/s004150050387

Vogel, A., Gade, A., Stokholm, J., and Waldemar, G. (2005). Semantic memory impairment in the earliest phases of Alzheimer's disease. *Dement. Geriatr. Cogn. Disord*. 19, 75–81. doi: 10.1159/000082352

Walenski, M., Sosta, K., Cappa, S., and Ullman, M. T. (2009). Deficits on irregular verbal morphology in Italian-speaking alzheimer's disease patients. *Neuropsychologia* 47, 1245–1255. doi: 10.1016/j.neuropsychologia.2008.12.038

Whatmough, C., and Chertkow, H. (2002). Category-specific recognition impairments in Alzheimer's disease. *Category Spec. Brain Mind* 181–210.

Wilson, R. S., Leurgans, S. E., Boyle, P. A., and Bennett, D. A. (2011). Cognitive decline in prodromal alzheimer disease and mild cognitive impairment. *Arch. Neurol.* 68, 351–356. doi: 10.1001/archneurol.2011.31

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer, VF, declared to the editor a co-authorship with one of the author GN, which occurred after their participation in this review process had been completed.

frontiers in Communication

ORIGINAL RESEARCH
published: 05 May 2021
doi: 10.3389/fcomm.2021.646454

Textual Effects in Compound Processing: A Window on Words in the World

Gary Libben[1], Jordan Gallant[1] and Wolfgang U. Dressler[2]*

[1]*Department of Applied Linguistics, Brock University, St. Catharines, ON, Canada,* [2]*Department of Linguistics, University of Vienna, Vienna, Austria*

We sought to move beyond single word and sentence processing experiments in order to examine textual effects on the processing of compound words in English. We developed minimal texts (sentences pairs that together constitute a story) that had neutral, semantic or lexical relations between the last word of the first sentence and the second word of the second sentence (which was always a compound noun). This generated minimal text triplets that differed only in the last word of the first sentence (e.g., "*She walked down to the path/river/water. The waterfall roared in the distance*"). Four experiments were conducted with a total 143 native speakers of English. Experiment 1 employed a Modified Maze Task to identify cross-sentence effects on compound processing. Sentence pairs with lexical links differed from those with semantic links, which, in turn differed from neutral pairs, providing evidence of cross-sentence influence on compound processing. In Experiments 2a, 2b, and 2c, we examined compound production using typing tasks. Results indicated that morphological effects found in single word typing persisted in text typing. In addition, constituent priming effects on typing were seen in both single word typing and sentence typing. Finally, morphological effects were correlated with overall story ratings. We thus conclude that morphological effects are not restricted to single word processing, but rather reflect the dynamics of real-world language processing.

Keywords: text, cohesion, coherence, compounds, morphology, priming, typing, maze task

OPEN ACCESS

Edited by:
Carlo Semenza,
University of Padua, Italy

Reviewed by:
Christina Manouilidou,
University of Ljubljana, Slovenia
Valantis Fyndanis,
Cyprus University of Technology, Cyprus

****Correspondence:***
Gary Libben
glibben@brocku.ca

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *26 December 2020*
Accepted: *15 February 2021*
Published: *05 May 2021*

Citation:
Libben G, Gallant J and Dressler WU (2021) Textual Effects in Compound Processing: A Window on Words in the World.
Front. Commun. 6:646454.
doi: 10.3389/fcomm.2021.646454

INTRODUCTION

A key challenge in the design of psycholinguistic research on lexical processing is to create experiments that have ecological validity and at the same time are sufficiently controlled so that specific variables and hypotheses regarding their effects can be examined. The achievement of ecological validity in lexical processing research can often be seen as the extent to which we are able to generalize experiment results to the processing of *words in the world*. We report on a series of experiments that have been designed to enhance ecological validity in the investigation of lexical processing by examining the processing of compound words within written texts. These texts have been constructed to more closely approximate naturally occurring language and at the same time be sufficiently controlled in their structure so that we can capture the interplay of factors in morphological processing. The specific structures that we have created can be described as minimal texts – two sentences that together constitute a story. An example of such a story is provided in 1) below:

1. "*She walked down to the path. The waterfall roared in the distance*".

In this minimal text story, the second word of the second sentence is the compound "*waterfall.*" A great deal of psycholinguistic research on the processing of words in isolation has found that the internal morphological structure of such words plays a role in how they are processed (Libben et al., 2020). Our goal in this research was to trace the processing of compound words such as *waterfall* from its characteristics when presented in isolation to its characteristics when presented in such minimal texts. We thus ask the question: Is the morphological processing found in single word processing also found in text processing?

In text linguistics and corresponding psycholinguistic studies, text comprehension has been shown to depend on textual (more precisely, co-textual) top-down-processing (Kintsch and van Dijk, 1978; Beaugrande and Dressler, 1981; Dederding, 1983; Ehrlich, 1991; Graesser et al., 1997; Kintsch, 1998; Verhoeven and Perfetti, 2008). From the first sentence of a text onwards, expectations are created in the hearer/reader about what is to follow, and these co-textual expectations are then corrected and elaborated by the on-going text. Hearers and readers use background knowledge to make predictions about what is going to happen next. This background knowledge has various sources. These include world knowledge, the situational context, and the preceding co-text. Expectations about what will be heard or read next constitute inferences (cf. Graesser et al., 1994; Clinton et al., 2015) that may affect the relative contribution of morphological processing during online processing. To the best of our knowledge, there exists a single paper (Smith et al., 2014) on the influence of situational context, but no studies on the impact of co-text on morphological processing.

Textual coherence is either established by the semantic and pragmatic means of relating text elements one to another or most directly via cohesive elements which bind text elements together on the textual surface, such as anaphoric pronouns or the definite article of a noun phrase which refer to their antecedents, or lexical repetition which, by default, signals anaphorically semantic identity or vicinity (unless differentiated through words, such as *another, a different*, etc., cf. Beaugrande and Dressler, 1981). In order to study textual anaphora (i.e., relations of an element to a preceding antecedent) one needs as minimal textual unit two coherent (and perhaps also cohesive) sentences. Thus, to take a specific example, if the first sentence of such a text is "*She walked down to the river.,*" textual coherence can be enhanced if the next sentence begins with an anaphoric pronoun (*e.g.,* "*Her feet felt…*"), or repetition (e.g., *Walking always helped…*), or a definite article (e.g., *The waterfall roared …*). This last type was the one employed in all our two-sentence texts.

A great deal of the literature on morphological effects in the processing of compound words such as *waterfall* has employed the priming technique (Forster and Davis, 1984) to address the question of whether the individual constituents *water* and *fall* are activated in the processing of the compound *waterfall*. The data, over a relatively large number of studies (*see* Sandra, 2019) provide evidence that this is the case – compound processing is facilitated by the prior presentation of one of its constituents (e.g., *water→waterfall*). In this research, we sought to track such priming effects from the domain of single word processing to the domain of minimal texts. Consider again, the two sentences in 1) above. Only by imagining pragmatically a bridging inference, might one suppose that the *path* might lead close to a *waterfall*. Would the story be made more coherent and would the processing of words within it be affected if the last word of the first sentence (i.e., *path*) were changed to a word such as *river* that bears a lexical semantic relation to the compound word *waterfall*? Such a case is shown in 2) below.

2. "*She walked down to the river. The waterfall roared in the distance.*"

And now, finally, would coherence and text processing be affected if the last word of the first sentence were changed so that it corresponded exactly to one of the constituents of the compound word *waterfall* as in 3) below? In this way, in addition to semantic coherence, the two-sentence text could also be said to possess cohesion.

3. "*She walked down to the water. The waterfall roared in the distance.*"

The research we report contrasts minimal texts such as 1), 2) and 3) in order to probe whether semantic and constituent priming effects that have been found in single word processing area also evident in the processing of minimal text. In this way, we are reporting a means to enhance the ecological validity of studies of processing and also to test the "ecological extendibility" of morphological effects in the processing of compound words.

Research on morphological processing in general and on compound processing in particular has suggested that lexical activation in reading involves automatic and obligatory access to the morphological constituents of words. The exact mechanism underlying this access is, to a large extent, still under debate (*see* Libben et al., 2020). However, despite the lack of agreement on whether constituent activation precedes whole-word recognition (Taft and Forster, 1976), follows it (Giraudo and Grainer, 2001) or occurs in parallel with it, (Baayen and Schreuder, 1999), there is a general consensus that morphological processing is not under strategic control by participants. Libben (2006), Libben (2014) argued that the morphological system is intrinsically organized to maximize opportunity for meaning activation (*see* Kuperman et al., 2009) and that this drive to maximize meaning opportunity results in obligatory activation of both whole-words and constituents.

It is important to note that in all the research referred to above, participants are focused on individual words in a manner that is unlikely to be representative of their normal real-world lexical processing (*see* Pollatsek et al., 2000). Visual lexical processing most often occurs in sentential and textual or discourse co-text and situational context. And, in such environments, any morphological processing of a single word would need to take its place among the many other processes and levels of processing that are active during text comprehension. Recently, studies have been reported that have attempted to bridge single word processing and sentence processing contexts. Mousikou and Schroeder (2019) found morphological effects using masked priming with affixed German words in isolation and also in sentence reading using the fast priming technique in eye tracking. In addition, Huang, et al. (2020), found a cross-modal constituent priming in a study in which Chinese compound words were heard in sentence context and visual lexical decision targets consisted of monomorphemic associates.

Studies such as these underline the advantages of linking the domains of single word processing and sentence processing. As we have indicated above, our goal in the present study is to go beyond the sentence level and to advance research that spans (and indeed bridges) the domains of single word processing and text processing by focusing on compound words in two-sentence texts such as those shown in 1), 2), and 3) above. This approach makes it possible to probe key features of morphological processing in a manner that both retains experimental control and enhances ecological validity. It also allows us to investigate whether morphological effects that have been observed in the study of single word processing could perhaps be artifacts of those types of studies. Indeed, it is critical that this possibility be investigated. In a single word processing experiment, there is, by definition, a single target word to process. It is therefore possible that the morphological effects that have been obtained do not generalize well to more natural language processing situations in which the processing of a single word is just one of many psycholinguistic activities being carried out. However, if these effects do generalize, we would have evidence that phenomena such as morphological constituent priming and morphological parsing play a role in the processing of text and, importantly also in the production of text (in which the words to be produced are part of the execution of a production plan (as opposed to an immediate early response to a surprise stimulus).

In the sections below, we describe the principles and procedures that we employed in order to bridge the investigation of lexical processing in single word experiments and lexical processing within texts. As we detail below, the present study focuses on English bimorphemic compound words. In our view, compound words represent an ideal stimulus type for an investigation such as this. The reason for this is that compounds are composed of identifiable constituents and those constituents can correspond to free-standing words.

STIMULUS DESIGN

Core Compound Stimuli

As pointed out by Dressler (2006) and Bauer (2016), compounds are core to word formation across languages. They thus provide a relatively simple and stable baseline from which to assess textual effects in lexical processing. A number of studies have shown that compound constituents are routinely activated in compound word processing (Libben 2014). There is also a relatively stable set of patterns that can be tracked across domains. One of these concerns the differences between morphological modifiers and heads. It is this difference that results in a compound such as *houseboat* being interpreted as a type of *boat*, but *boathouse* being interpreted as a type of *house*. In English, as a default, the morphological head is always the final element of the compound (with the exception of some unproductive patterns, e.g., *daredevil*, *passport*). In other languages however (e.g., Hebrew, Italian) it can be the initial element (Goral et al., 2008; Marelli et al., 2009).

Another matter that has been dominant in the compound processing literature concerns the extent to which a compound word is semantically transparent (e.g., Sandra, 1990; Zwitserlood, 1994; Ji, Gagné and Spalding, 2011; Davis et al., 2016). These factors have been reported to interact in English (Monahan et al., 2008; Fiorentino and Fund-Reznicek, 2009; Gagné and Spalding, 2009; Ji et al., 2011; Gagné and Spalding, 2014). This interaction was investigated in a compound typing task reported by Libben and Weber (2014). They classified compound words in terms of the extent to which the meaning of each constituent within the compound corresponds to its meaning as an independent word. Thus, a compound such as *bedroom* was classified as transparent-transparent (TT), a compound such as *grapefruit* was classified as opaque-transparent, a compound such as *jailbird* was classified as transparent-opaque, and finally a compound such as *humbug*, for which the meaning of neither constituent plays a role in the meaning of the compound as a whole, was classified as opaque-opaque (OO). They found that all four compound types showed elevated typing times at the compound constituent boundary. However, this boundary effect was attenuated for the OO compounds in comparison to the TT, OT, and TO compounds, which patterned together.

Libben (2014) has argued that in order to understand the dynamics of compound processing, it is important to distinguish between a word such as *water* in isolation and its homographic counterpart as a constituent *water*-in the compound *waterfall*. The reason for this is that a lexical priming experiment may, in fact, place these representations in competition. In an isolated word constituent priming experiment, however, it is not clear whether a constituent prime is perceived as a separate word, or a foreshadowing of a constituent. However, in the types of stimuli that we have employed in this study [e.g., sentences 1), 2) and 3) above], putative primes are presented as part of the text preceding a target compound word. Thus, there can be little debate over whether the prime is functioning as an independent word. In this way, the paradigm we employ, namely one in which compound processing is studied in sentential context creates a special opportunity to disentangle the potential facilitatory effects of repetition and the potentially inhibitory effects of competition.

The Creation of Two-Sentence Texts to Investigate Lexical Processing in Context

In the experimental paradigm reported in this paper, we developed balanced and structured two-sentence texts in order to explore compound processing in context. These two-sentence texts constituted the core stimuli in the study. Each of the two sentences was six words in length. Together, the sentences were designed to represent a coherent and cohesive (either minimally or very cohesive) text through which the interaction between

TABLE 1 | A stimulus triplet. Each member of the triplet is composed of two sentences of six words each.

Prime Category	Sentence Pair
Neutral	*She walked down to the path. The waterfall roared in the distance*
Semantic	*She walked down to the river. The waterfall roared in the distance*
Lexical	*She walked down to the water. The waterfall roared in the distance*

lexical and textual effects could be investigated. A minimal text which consists only of two sentences has the advantage of excluding the influence of preceding but non-adjacent sentences.

To examine lexical influences in text, we ensured that, for each of our 56 six-word-sentence pairs, the first sentence of the pair served to establish the set of expectations that might affect the processing of the second sentence and, in particular, the target compound word within that second sentence. For all sentence pairs, the target compound word was always the second word of the second sentence. In all sentence pairs, this target word was immediately preceded by the word "*The*", which, as a definite article, is itself a cohesive element. The definite article (which constituted the first word of every second sentence in the pair) was immediately preceded by one of the three types of antecedent conditions (neutral, semantically-related, lexical compound constituent).

These three conditions map onto the sentence pair triplet shown in **Table 1**. Together this triplet constitutes a within-item manipulation in which a single compound word (in this case, *waterfall*) is preceded by a sentence in which the final word completes the textual baseline control condition in which the first sentence of the text is textually (i.e., text-semantically) consistent with the second sentence of the text, but does not cue the target compound word explicitly. We therefore label this as the neutral condition. This is contrasted to the semantic condition in which the final word of the first sentence is a semantic associate of the target compound (in this case *river* - > *waterfall*) and thus creates an additional link of coherence. Finally, the lexical condition is one in which the last word of the first sentence corresponds to a constituent of the compound (in this case, *water* - > *waterfall*) and thus represents, in addition, lexical repetition as a means of strengthening cohesiveness.

This stimulus structure shown in **Table 1** constitutes the foundation of our investigation. By using this central structure, we sought to investigate morphological processing against the background of a common two-sentence minimal text structure. The common stimulus structure (two sentences of six words each) was designed to limit background variability in the experiment. The fixed positions of the compounds enabled us to ensure that they played a similar syntactic role in the sentence. Finally, the fixed position of the prime allowed us to standardize the distance of the prime to the target and to the sentence boundary, thus reducing textual and syntactic variability. Against this background, we could thus bring into the experiment, the key factors of transparency and morphological role (head vs. non-head) which, as we have noted above, have played an important role in the psycholinguistic literature on the online processing of compound words. To examine these effects, our stimuli were created so that half of the target stimuli were compounds containing two transparent (T-T) constituent morphemes (n = 28). The other half were partially transparent compounds, containing one semantically transparent and one opaque constituent. Of these, the modifier constituent was transparent (T-O) for 14 stimuli and the head was transparent (O-T) for the other 14 stimuli.

All prime words had the same lexical category within the first sentence of the text and were selected to be as comparable as possible in terms of lexical frequency. Lexical primes corresponded to the structures used in traditional constituent priming experiments in which a compound constituent serves as the prime and the entire compound serves as the target. However, because our design required that this priming relation be plausible in a two-sentence text, only semantic transparent constituents could be involved. The reason for this is that it would not be possible to create a coherent text in which the opaque constituent serves as the last word of the first sentence. Consider, for example, the first text with an OT compound shown in the **Supplementary Material**:

4. "He called everyday about the road/crack/hole. The pothole still hadn't been fixed."

Although all three priming alternatives create a coherent text, the use of the opaque first constituent (*pot*) would not. Rather it would result in a situation in which either the text is noncoherent or the compound meaning would need to be changed: "He called everyday about the pot. The pothole still hadn't been fixed."

For semantic primes, our design requirement that all priming conditions result in a coherent two-sentence text necessitated the use of semantic associates that were related to the compound as a whole, rather than any particular constituent. This can also be seen in the relation between the prime *crack* and the compound *pothole* in the example above.

Finally, the design resulted in the neutral primes being more related to the compound target than it would typically need to be in a single word priming experiment. The reason for this, again, is that it must result in a coherent two-sentence text (e.g., "*He called everyday about the road. The pothole still hadn't been fixed*"). The resulting priming design is summarized in **Table 2**.

The stimulus design considerations discussed above and summarized in **Tables 1, 2**, enabled the development of the four experiments described below. These experiments call into two groups: Experiment 1 uses the Maze Task to examine priming effects in the reading of compounds in texts. Experiments 2a, 2b, and 2c all utilize the typing task as a window into within-word processing. Experiment 2a acquires baseline data for keystroke latencies in the typing of compound words. Experiment 2b adds a constituent priming paradigm to that typing paradigm, and Experiment 2c embeds primes and compound targets in

TABLE 2 | The three categories of compound stimuli and their associated prime words.

Category[a]	Example Compound	Neutral Prime	Semantic Prime	Lexical Prime 1	Lexical Prime 2	Distractor (Exper. 1)
TT1	waterfall	path	river	water	N/A	alfalfa
TT2	teacup	set	China	N/A	cup	alfalfa
OT	pothole	road	crack	N/A	hole	avarice
TO	bookworm	series	novel	book	N/A	psyllium

[a]*Note: T = transparent; O = opaque; TT1 = transparent-transparent compound primed by first constituent; TT2 = transparent-transparent compound primed by second constituent.*

two-sentence minimal texts. Together, these experiments are designed to address the following questions:

Question (a): Are priming effects found in the processing of individual compound words also evident when the compound is part of a text and the putative prime is an antecedent word within that text?

Question (b): Can the typing paradigm be used to examine constituent priming effects in the processing of compound words?

Question (c): Are the elevated keystroke latencies found at the morphological boundary in the typing of single compound words also evident when the compound is part of a text?

By addressing Question (a), we hope to assess the extent to which we can relate the highly constrained designs of lexical and morphological priming within single-word experiments to the links that can naturally occur among words within texts.

By addressing Question (b), we examine whether there is opportunity to use the typing task to both tap lexical activation (which is often revealed through priming effects) and lexical production (which we expect to be less driven by the characteristics of stimulus presentation). Moreover, if the typing task can be used in this way, it makes a rich set of dependent variables, linked to particular locations within a word, available for analysis.

Finally, by addressing Question (c), we seek to determine whether the morphological effects that have been linked to the typing of words in isolation are also found when words are typed in more natural texts. If this is the case, it would increase our confidence that these effects are relevant to the processing of "words in the world."

EXPERIMENT 1: COMPOUND READING IN A MODIFIED MAZE TASK

In this experiment, we investigated whether compound processing is influenced by lexical semantic association (e.g., river → waterfall) and additional lexical constituent overlap (e.g., water→ waterfall) across sentences within minimal texts.

The experiment focused on reading times within a sentence. Building upon the report of Irsa et al. (2016) and the observations of Gallant and Libben (2020), we employed a Modified Maze Task. In this task, participants read text in a word-by-word manner as in a self-paced reading task. As noted by Forster et al. (2009), however, self-paced reading can lead participants to develop a reading rhythm. Because the focus of our study was a particular location (the compound word) in the text, we employed the maze task (Forster et al., 2009; Forster 2010). The task offers methodological advantages in the investigation of self-paced reading, particularly in cases in which researchers are targeting incremental effects.

As in self-paced reading, sentences in the maze task are presented word-by-word. However, each target word is presented alongside a distractor item. Participants must select the target to continue "weaving their way" through the sentence. Thus, each decision juncture has the effect of a "speedbump," forcing participants to incrementally integrate each successive target into the sentence context and facilitating the observation of highly processing.

Gallant and Libben (2020) argue that lexical choice junctures need not be placed at every sentence position in order to observe localized lexical effects during reading. Rather, by selectively introducing lexical choices at key sentence positions, such effects can be observed without introducing additional processing load and the potential for spurious pre-activation caused by the introduction of numerous distractor items. In Experiment 1, it was this simplification of the maze task that was employed.

As is detailed below, the maze task enables a focused examination of the amount of time to choose a compound word (e.g. *waterfall*) over a distractor word (e.g., *alfalfa*) in sentence reading. We hypothesized that, if an inter-sentential "priming effect" were to be obtained, then we would see lower decision response times for the lexical and semantic conditions, as compared to the neutral condition. Such a finding would constitute an affirmative answer to the question of whether priming effects found in the processing of individual compound words are also evident when the compound is part of a text and the putative prime is an antecedent word within that text.

METHOD

Participants

Twenty-three native speakers of English (13 female, 10 male) residing in the United States participated in the experiment. Participants were recruited to the experiment using Amazon's Mechanical Turk crowdsourcing platform. Access to the experiment was restricted such that only Mechanical Turk workers with experience (>100 tasks complete) and a high approval rating (>80%) were eligible to participate. All

TABLE 3 | Example priming conditions for one sentence-pair stimulus including the target and matched distractor (target = *waterfall*, distractor = *alfalfa*).

Prime Category	First Sentence	Second Sentence
Neutral	She walked down to the path	The waterfall alfalfa roared in the distance
Semantic	She walked down to the river	The waterfall alfalfa roared in the distance
Lexical	She walked down to the water	The waterfall alfalfa roared in the distance

TABLE 4 | Linear Mixed Effect model of lexical choice latency in the Experiment 1 maze task (neutral prime condition is on the intercept).

Fixed Effect	Estimate	Std. Err	df	*t*-value	*p*-value
(Intercept)	1,664.81	61.50	45	27.07	<0.001
Prime Category: Lexical	−168.22	19.95	908	−8.43	<0.001
Prime Category: Semantic	−93.87	20.57	911	−4.56	<0.001
Compound Frequency (log)	−18.09	4.7	65	−3.85	<0.001
Trial Order	−3.60	0.5	921	−7.16	<0.001

participants were required to have shown a response accuracy above chance (50%). The age of participants ranged from 23 to 69 with a mean of 37 (SD = 12). All participants had completed high school and 18 had received some post-secondary education. No visual or motor-articulatory impairments were reported by any participants.

Materials

The stimuli used in Experiment 1 were the 56 sentence pairs described above and 56 monomorphemic lexical items used as distractors in the modified maze task. In the design of our modified maze task, it was important that distractors be grammatically acceptable (i.e., nouns), but that they not be semantically and pragmatically plausible continuations of the text. This consideration ensured that response latencies would have a greater chance of reflecting prime-target properties. Accordingly, as a measure of experiment control, all three priming versions of a two-sentence text were linked to a single distractor item. In order to ensure the comparability of distractors across the set of two-sentence texts, the following materials development procedure was employed: The English Lexicon Project database (Balota et al., 2004) was used to create a list of potential distractor words that were comparable to the compound stimuli in terms of number of letters, mean bigram frequency, and wordform frequency. A subset of 56 distractors were randomly selected from this initial list, and those 56 distractors were randomly assigned to the 56 two-sentence texts. An example result of this procedure is shown in **Table 3**, which contains the stimulus triplet for the compound *waterfall* and its assigned distractor item (*alfalfa*).

Procedure

The experiment was created in PsychoPy3 (Peirce et al., 2019) and hosted on the Pavlovia open behavioral-science platform. Participants accessed the experiment via a URL posted on Mechanical Turk. Each session began with a five-item demographic questionnaire asking for participants Mechanical Turk Worker ID, age, educational achievement, and native language. The final question asked participants whether they had any visual or motor-articulatory impairments.

Using the Modified Maze method, a single lexical choice juncture was placed at the target compound position (i.e., the second word of the second sentence). In each trial, participants read the sentence pairs word-by-word. The task began with a fixation point in the center of the screen, after which, the targets presented. Participants used the "up" arrow key to indicate that they were ready to proceed to the next target. This procedure continued until participants reached the lexical choice task, where the target compound and distractor appeared together in the middle of the screen separated by a ~10 cm gap. The participants then used the "right" and "left" arrow keys to select the item they felt best fit the sentence context. Once they made their selection, the procedure continued as normal until the end of the second sentence. The response latency of each arrow key press was recorded. Trials in which the distractor was chosen instead of the target were marked as error trials. The order of trials was re-randomized for each participant.

Immediately after reading each sentence pair, participants were asked to rate its coherence on a 5-point scale. Participants were shown the whole sentence pair again and were provided with a description of both extremes on the scale: "1 = Together, the two sentences do not seem to make a story at all" and "5 = Together the two sentences make a connected, believable and interesting story." Participants indicated their rating by clicking on a rating scale. They were able to change their response prior to submitting it by clicking on a different position. Once they were satisfied with their rating, they submitted it using the spacebar (at which time all selected ratings and response times were recorded).

RESULTS

Trials with lexical choice RTs of greater than 2,000 ms ($n = 263$) were removed to ensure that the data reflected on-line lexical processing. All error trials were also removed ($n = 79$). The final subset of data used in the analysis contained 946 correct lexical choice responses from 23 participants.

A linear mixed effects model was created to determine whether the preceding co-text influenced lexical choice RT. The key variable included in this model is prime category, which represents the varying degree of textual coherence across sentence pairs in each stimulus triplet. Participant and compound noun were included as random effects and trial order and compound frequency were included as control variables. A summary of the model is provided in **Table 4**.

As is shown in **Figure 1**, model estimates indicated that lexical choices were significantly faster in both the semantic and lexical prime categories. Compared to the neutral condition, lexical choice RT was estimated to be 168 ms faster in the lexical

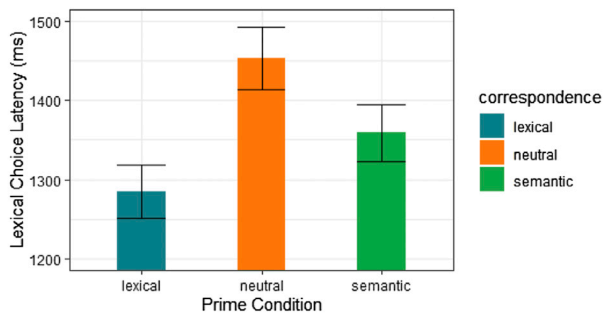


FIGURE 1 | Model estimates of lexical choice latency (in milliseconds) to correctly choice a compound over a distractor in the modified maze task under three text priming conditions (lexical, neutral, semantic).

prime condition ($t = -8.43$, $p < 0.001$) and 94 ms faster in the semantic prime condition ($t = -4.56$, $p < 0.001$). Using the "emmeans" package in *R* (Lenth, 2020), we obtained least square means for each prime condition and then used the "contrast ()" function to compare each level. There were significant differences among all levels ($p < 0.0001$).

We thus observed that lexical choice was facilitated by lexical semantic association (e.g., *river* → *waterfall*) and additional lexical constituent overlap (e.g., *water*→ *waterfall*) between the last word of sentence 1 and the second word of sentence 2. We did not observe an effect of semantic transparency or an interaction of semantic transparency and prime condition. This variable was therefore not included in the model shown in **Table 4** (which also improved model fit). Thus, for all compound types (TT, OT, and TO), recognition was affected by prior presentation of a transparent constituent. This finding is consistent with the view that the presence of semantic opacity in the compounds does not diminish the extent to which it is processed as a morphologically structured word (Libben, 2014). Finally, we interpret these results to show that lexical priming effects are revealed within the modified maze task paradigm and that these effects can span sentence boundaries within text.

EXPERIMENT 2A: COMPOUND WORD TYPING

In Experiment 1 reported above, we found evidence that compound processing is affected by lexical semantic association and lexical constituent overlap across sentences within a minimal text.

In Experiments 2a, 2b, and 2c we use a typing production paradigm that enables us to build upon Experiment 1 in order to look inside the processing of compound words and thus address our second and third questions:

Question (b): Can the typing paradigm be used to examine constituent priming effects in the processing of compound words?

Question (c): Are the elevated keystroke latencies found at the morphological boundary in the typing of single compound words also evident when the compound is part of a text?

The Typing Task

Experiments 2a, 2b, and 2c all involve the examination of lexical production effects as revealed by the typing task. As has been shown in a number of recent investigations (e.g., Weingarten et al., 2004; Libben and Weber, 2014; Gagné and Spalding, 2016; Libben et al., 2016), typing presents an effective window into the processes of lexical production. As a natural and well-practiced production activity it has high ecological validity. Importantly, typing enables us to focus on particular locations within a word in a straightforward and effective manner.

It is important at the outset to underline the fact that typing is a language production activity. Thus, in contrast to activities such as lexical decision, which can be seen as providing a measure of lexical access, the typing task tracks the unfolding of language production over time. This is particularly relevant to an understanding of the processing of compound words, which are the focus of the present study. On average, it has been found that the typing of a letter in a compound word takes about 200–300 ms (Libben and Weber, 2014). Thus, the full production of an eight-letter compound word will likely take more than 1,600 ms. This is at least twice the amount of time that is normally seen in lexical decision latencies to compound words. During that period of processing time, the typing paradigm offers a window into ongoing cognitive activity and its potential relation to the characteristics of a word and those of its sub-elements. It has been shown that lexical constituent boundaries within compounds emerge as identifiable structures within compound typing because they are associated with larger keystroke intervals. In other words, participants would take more time to type the medial "f" in the compound *waterfall*, than they would take to type surrounding letters (Sahel et al., 2008; Libben and Weber, 2014; Gagné and Spalding, 2016). Thus, these studies have linked compound morphological structure to online language production. Under such a view, compound constituents can be considered to be units of planning and execution, so that the typing of a compound such as *waterfall* is carried out as the typing of *water, pausing,* and then the typing of *fall*. In Experiments, 2a, 2b, and 2c, we investigate whether this correspondence between keystroke latencies and compound morphological structure can be affected by lexical priming and by the text environment in which the compound is typed.

In all three experiments, our primary focus was the typing of the compound words that constituted the second word of the second sentence in each sentence pair (e.g., *waterfall* in the example shown in **Table 1**). In Experiment 2a, we examined the typing of the compound word in isolation. This allowed us to assess whether morphological constituent boundary effects are seen for these words when they are presented in isolation. In Experiment 2b, we built on this by examining priming effects in the single-word typing of these stimuli. To the best of our knowledge, this is the first reported study of constituent priming in compound word typing. Finally, in Experiment 2c, we were in a position to build on the single-word typing experiments to examine whether between-sentence priming effects are seen and to examine whether there are effects of morphological structure when compounds are typed in sentence context and how typing patterns may relate to participants' ratings of the extent to which the sentence pairs form a story.

METHOD

Participants

Thirty-one participants (female = 19, male = 12) were recruited using Mechanical Turk. The same selection criteria as Experiment 1 were applied. All participants were native speakers of English residing in the United States. The age of participants ranged from 21 to 61 with a mean of 36 (SD = 10). The majority of participants were university graduates (n = 17). The remaining participants had either completed high school ($n = 4$), an MA ($n = 5$), or some college courses ($n = 5$). No participants reported any visual or motor-articulatory impairments.

Materials

The stimuli in this experiment were the 56 compound words that constituted the core of the study. All compounds were English nouns and were composed of two constituents. The compound words ranged in length from seven to twelve letters (mean = 8.4). The set of 56 compounds contained 28 fully transparent compounds and 28 that contained an opaque constituent (14 Opaque-Transparent and 14 Transparent-Opaque).

Procedure

This experiment was designed to collect baseline compound typing data against which we could compare the primed compound typing data from Experiment 2b and the compound typing within texts in Experiment 2c. Accordingly, the task was designed to contain as little experimental manipulation as possible. From the perspective of the participants, this experiment was a simple copying task. A compound word appeared in the middle of the computer screen and they were asked to type it immediately below. The stimulus remained on the screen during the typing process.

Each session began with the same 5-item demographic questionnaire from Experiment 1. There were 60 trials in total: four practice trials and 56 experiment trials. Thus, each compound was seen once and typed once. The order of compounds was in a different random order for each participant.

Each trial began with a 1,000 ms fixation point in the middle of the screen, after which the target word was presented. A marker (">>>") also appeared below the target, indicating the position on the screen where the participants typed input would appear. Participants copied out the word using a computer keyboard. When they had finished typing the word, they pressed the "return" key. This initiated the start of the next trial. Prior to pressing the "return key," participants were able to correct typing errors using the "backspace" key. All keystrokes and their corresponding latencies were recorded.

RESULTS

Data Preparation

To begin, trials containing typing errors were removed ($n = 275$). This included trials in which typing errors were made and subsequently corrected. One participant, with 0% typing accuracy, was removed as a result. To ensure that typed responses reflected on-line, automatic processing, trials containing inter-keystroke intervals (IKIs) greater than 2,000 ms ($n = 91$) were removed. Participants' mean inter-keystroke intervals were analyzed to identify any potential outliers in terms of typing ability. A cut-off was set at three standard deviations from the group mean. However, no outlier participants were identified. After data trimming, the final data set used in the analysis contained 11,236 individual keystroke responses from 1,370 out of 1,736 trials (22% item removal). This rate accorded with previous patterns in the analysis of typing data—error rates tend to be high because any error in any part of the string requires its removal (because any error can affect typing times at other parts of the string).

Modeling of Within-Word Typing

Linear mixed effects modeling was used to analyze compound typing. Participant, compound word, and individual letter were included as random factors. Individual letter was included to capture variance originating from the arrangement of keys on the keyboard and the idiosyncrasies of participants' typing styles and hand positions.

Our analysis focused on five letters within a compound word target—the two letters preceding the constituent boundary, the first letter of the second constituent (taken to be the constituent boundary position), and the two following letters. Thus, for the compound word *waterfall*, these five letters would be *e, r, f, a, l*, where the time taken to type the letter "f" was seen as the time taken at the constituent boundary. This was the key variable of analysis, allowing us to test whether there was elevation of typing times at the constituent boundary, as had been reported by Libben and Weber (2014) as well as Gagné and Spalding (2016).

The linear mixed effects model is shown in **Table 5**. As can be seen in this table, the model included the lexical frequencies of the free-standing words that corresponded to the compound modifier and compound head (e.g., *water* and *fall* for the compound *waterfall*). As is shown in the model, higher constituent lexical frequencies for the head constituent were associated with lower typing times. The effect of lexical frequency of the modifier did not reach significance ($p = 0.07$). We saw lower typing times associated with compounds that had higher modifier positional family sizes (i.e., the number of different compounds that begin with a particular constituent). In contrast, typing within compounds with longer heads was slower. Finally, as expected, the control variable, trigram frequency, was associated with lower typing times. A trigram consisted of the letter being typed and the letters immediately preceding and succeeding it. As a control variable, it captured the automaticity developed through repeated production of letter combinations.

Neither whole-word frequency nor semantic transparency showed significant effects and were, therefore, not included in the model, improving model fit.

Results provide evidence for morphological processes in the planning and execution of compound word typing. As shown in **Figure 2**, a significant elevation in IKI latency was observed at the boundary position between morphological constituents. This

TABLE 5 | Linear Mixed Effects modeling of inter-keystroke interval (IKI) latency at positions relative to the morpheme boundary (position -2 is the intercept) for typing in isolation in Experiment 2a.

Fixed Effect	Estimate	Std. Err	df	*t*-value	*p*-value
(Intercept)	5.680	0.14	140	40.90	<0.001
Typing Position: −1	−0.110	0.02	5,581	−4.79	<0.001
Typing Position: Boundary	0.450	0.03	3,586	17.87	<0.001
Typing Position: +1	<0.001	0.02	6,467	−0.09	0.930
Typing Position: +2	−0.050	0.02	6,204	−2.57	0.010
Head Constituent Length	0.040	0.01	51	2.96	<0.001
Trigram Frequency (log)	−0.060	0.01	1763	−4.57	<0.001
Modifier Word Frequency	−0.040	0.02	49	−1.87	0.07
Modifier Positional Family Size	−0.003	0.001	50	−2.21	0.03
Head Word Frequency	−0.060	0.02	49	−3.81	<0.001

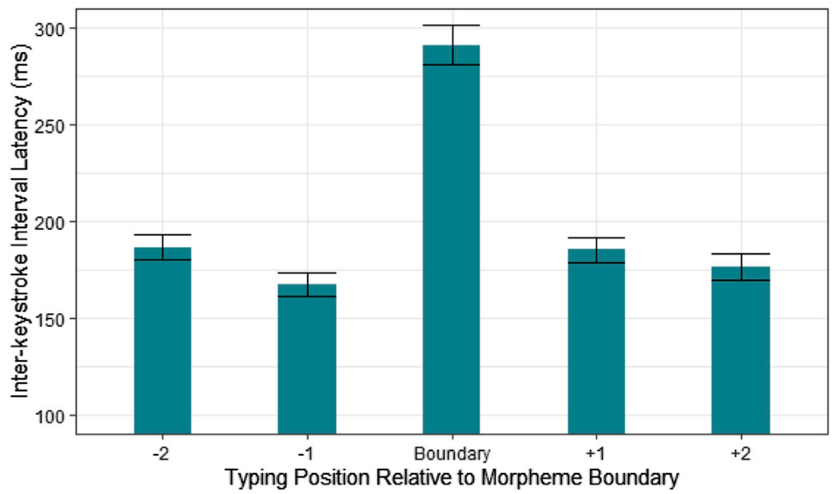


FIGURE 2 | Model estimates of logged inter-keystroke interval (IKI) latencies for typing positions relative to the morpheme boundary in Experiment 2a. The boundary position is the first letter of the second constituent. Two positions both before and after the boundary are shown.

shows that the morphological structure of a compound word provides an organizational framework by which compounds can be decomposed, planned, and produced using constituent level production units. The presence of constituent boundary effect in typing latencies across TT, OT, and TO compounds is consistent with the results of Libben and Weber (2014).

A significant facilitatory effect of modifier positional morphological family size was also observed ($t = -2.21$, $p = 0.03$), despite being highly colinear with modifier word frequency (−0.547). Following Wurm and Fisicaro (2014), we opted not to residualize these predictors, and instead, included both in the model. This result supports the notion that compound production is more fluent when the modifier constituent commonly occurs in that position.

EXPERIMENT 2B: PRIMED COMPOUND TYPING

In this experiment, we compared the typing of our core compound stimuli across three priming conditions: neutral, semantic, and lexical. The prime words used in this experiment were the same words that appear in the sentence conditions shown in **Table 1**. This was done to determine what types of priming effects are found when the compounds are produced as isolated words, thus setting the stage for our investigation of morphological effects in text typing in Experiment 2c. As we have stated above, to the best of our knowledge this is the first reported study of a constituent priming experiment which involves the typing of target stimuli. Thus, in addition to serving as a bridge to the typing of two-sentence texts in Experiment 2c, this experiment examined the extent to which constituent priming effects can be obtained when the dependent variable is keystroke latency within compounds.

METHOD

Participants

Twenty-nine participants (female = 4, male = 25) were recruited using Amazon Mechanical Turk. The same previous selection criteria were applied. All participants were native speakers of

English residing in the United States. The age of participants ranged from 23 to 63 with a mean age of 39 (SD = 12). The majority of participants were university graduates (n = 20). The remaining participants had either completed high school (n = 2), an MA (n = 5), or some college courses (n = 2). No participants reported any visual or motor-articulatory impairments.

Materials

The core visual stimuli for Experiment 2b included the same 56 core compound stimuli from Experiment 2a as well as the 168 prime words which appear in position six of the sentence pair stimuli used in Experiment 1 (i.e., the final word of the first sentence). Each of the 56 core stimuli appeared in each of three prime conditions: lexical, semantic, and neutral, which were counterbalanced across sessions.

Procedures

The procedure mirrored that of Experiment 2, except for the inclusion of an unmasked prime which was presented visually prior to the compound stimulus. All participants saw all compound targets once, in one of three priming conditions. Each trial began with a 2,000 ms fixation cross in the center of the screen. This was followed by the presentation of the prime word for 100 ms. Following this, the screen went blank for 300 ms buffer, after which the compound stimuli were presented. After this point, the procedure was identical to Experiment 2a—participants typed the compound word. The compound stimulus remained on the screen during typing and the trial was terminated when the participant pressed the "return" key.

RESULTS

Data Preparation

The same data preparation method was identical to that used in Experiment 2a. Trials containing typing errors (n = 261) and those containing IKI greater than 2,000 ms (n = 96) were removed. The resulting data set included 11,711 individual keystroke responses from 1,267 out of 1,624 trials. Thus, 357 trials (28%) were removed from the analysis.

Modeling

Linear Mixed Effect models were constructed for both typing onset latency and morpheme boundary IKI latency. As we have indicated above, the IKI at the constituent boundary during the typing of a compound word can be seen as reflecting the initiation of a motor plan for the typing of the second constituent of the compound. Interpreting the typing latency at the onset of typing, however, is somewhat more complex. The reason for this is that word-onset latency likely has a greater number of components. It has a forward-looking component because it involves the initiation of a motor plan for the typing of the first constituent (as well as the sequence of constituents for the word as a whole). In the paradigm that we have employed, however, it also has a retrospective component. Because typing can begin as soon as the compound appears on the screen, we expect that the time taken to initiate the typing of the compound will reflect the complexity of the processes involved in its recognition.

Typing Onset Latency

Our analysis of typing onset latency focused on three types of priming – neutral, semantic, and lexical. We also examined which of the two compound constituents was primed in the lexical condition. For TO compounds and the first set of TT compounds, this was always the first constituent. For OT compounds and the second set of TT compounds, this was always the second constituent. We also included a control variable, Trial Order, to capture practice effects over the course of the experiment. As can be seen in the linear mixed effects model in **Table 6** and in **Figure 3**, lexical priming to the modifier (i.e., the first) constituent of the compound resulted in faster typing onset, as compared to all other priming conditions. Thus, compound typing onset is most affected when the first constituent to be typed was presented visually as a prime word. As is also shown in **Table 6**, participants took less time to initiate typing as they progressed through the experiment.

Within Word Typing

As in Experiment 2a, the analysis of typing times within the compound word was focused on the first letter of the second constituent (as the constituent boundary), the two letters leading up to it, and the two letters following it. As can be seen in the linear mixed effects model shown in **Table 7**, typing times were elevated at the boundary position, thus replicating the pattern seen in Experiment 2a. The pattern of effects for the additional variables also accorded with the pattern seen in Experiment 2a. Head frequency, modifier positional family size, and trigram frequency increased typing speed; head length slowed it down. The results with respect to the effect of modifier frequency were aligned with those of Experiment 2a—they showed a trend toward facilitation, but failed to reach significance (p = 0.14). Finally, as in Experiment 2a, within word typing times were not affected by whole-word frequency or compound transparency.

EXPERIMENT 2C: SENTENCE TYPING

This experiment builds upon Experiment 1, 2a, and 2b by examining the production of the two sentence minimal texts used in Experiment 1 in a typing task. Our analysis focused on the typing of the compound word that occurred as the second word of the second sentence in the text. We examined whether the morphological effects that we observed in single-word typing are also seen when lexical production is part of the production of meaningful coherent text. This question is critical to our understanding of whether morphological processing could be an artifact of the attention that can be allocated to the internal structure of a word under conditions in which that word is presented in isolation.

We also examined lexical priming during the production of text. In a classical lexical priming paradigm, each trial contains only two stimuli—the prime and the target. Our goal in this

TABLE 6 | Linear Mixed Effects modeling of typing onset latency of the modifier compound constituent for compounds typed in isolation in Experiment 2b.

Fixed Effect	Estimate	Std. Err	df	*t*-value	*p*-value
(Intercept)	6.93	0.06	29	108.80	<0.001
Prime Condition: lexical - modifier constituent	−0.06	0.02	1,102	−3.72	<0.001
Prime Condition: lexical - head constituent	0.01	0.01	1,130	0.55	0.58
Prime Condition: semantic	0.001	0.01	1,135	0.05	0.96
Trial Order	−0.001	<0.001	1,139	−4.42	<0.001

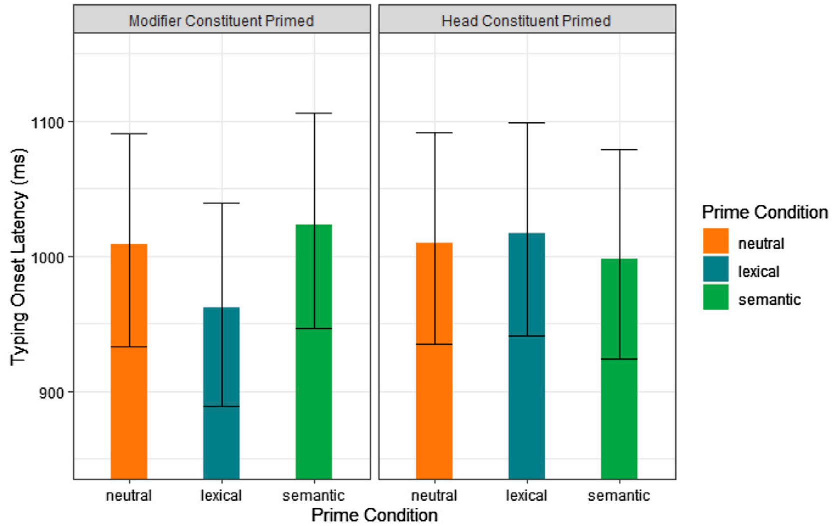


FIGURE 3 | Model estimates of typing onset latencies for prime conditions and the constituent primed for compounds typed in isolation in Experiment 2b.

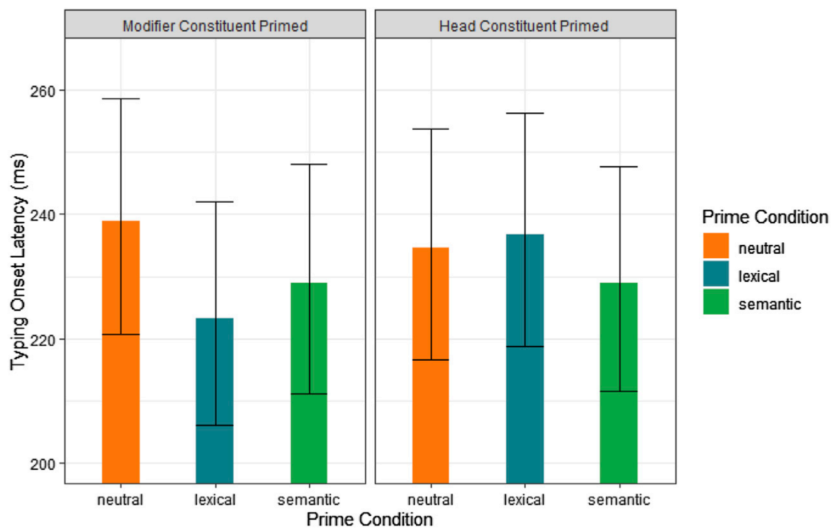


FIGURE 4 | Model estimates of typing onset latencies for prime conditions and the constituent primed for compounds typed in sentence context in Experiment 2c.

TABLE 7 | Linear Mixed Effects modeling of inter-keystroke interval latency at positions relative to the morpheme boundary (position -2 is the intercept) for compound typing in Experiment 2b.

Fixed Effect	Estimate	Std. Err	df	*t*-value	*p*-value
(Intercept)	5.70	0.14	160	42.24	<0.001
Typing Position: −1	−0.12	0.02	4,408	−5.34	<0.001
Typing Position: Boundary	0.41	0.03	1867	15.76	<0.001
Typing Position: +1	−0.01	0.02	5,546	−0.72	0.47
Typing Position: +2	−0.06	0.02	5,176	−2.62	0.01
Head Constituent Length	0.02	0.01	50.74	1.97	0.05
Trigram Frequency (log)	−0.05	0.01	1706	−3.76	<0.001
Modifier Word Frequency	−0.03	0.02	47.23	−1.50	0.14
Modifier Positional Family Size	−0.002	0.001	48.03	−2.02	0.05
Head Word Frequency	−0.07	0.02	49.92	−3.79	<0.001

experiment was to understand the extent to which the priming patterns observed in Experiment 2b would also be seen under more natural text processing conditions in which the prime target relations coexist with other within-text relations.

METHOD

Participants

Participants were recruited using Mechanical Turk with the same previous eligibility restrictions as were used for Experiments 1, 2a, and 2b. Data were collected from 61 (female = 33, male = 28) participants living in the United States. Participants ranged in age from 23 to 62, with an average of 40. One participant reported being near sighted. No other participants reported any visual impairments. Most participants were university graduates (n = 39). Other had completed high school (n = 8) or graduate school (n = 3). Some participants also indicated that they had completed some college courses (n = 9). Two participants did not respond to this question.

Procedure

Each session of Experiment 2c began with the same five-item demographic questionnaire used in previous experiments. Each trial began with a 2,000 ms buffer followed by the presentation of the target sentence pair as a single line of text centered in the middle of the screen. Thus, unlike Experiment 1, where sentence pairs were presented incrementally, here the entire pair was presented all at once. As participants typed out the sentence pair, their typed input appeared on the screen directly below the target next to a marker (“>>>”). The two-sentence text remained on the screen during the typing process. Each individual keystroke and its response latency relative to the beginning of the trial was recorded. Once participants had finished typing, they submitted their response by pressing the “return/enter” key. Participants then completed the same sentence pair rating task (as in Experiment 1) and went on to the next trial.

Materials

The same 56 core sentence pairs used in Experiment 1 were used in this experiment (*see* description in **Table 1** and full stimulus list in the **Supplementary Material**).

RESULTS

Data Preparation

Compound typing data were extracted from the overall data set and trimmed following the same criteria as were used for Experiment 1 and 2a. Trials containing typing errors prior to or during the typing of the compound word were removed (n = 622). Trials in which errors were made after the typing of the compound were retained. Correct trials containing IKI greater than 2,000 ms were also removed (n = 146). The final data set included 24,823 individual keystroke responses from 2,648 out of 3,416 trials (22.5% item loss).

Modelling of Typing Onset Latency

As in Experiment 2b, priming effects were examined in a linear mixed effects model using typing onset time as the dependent variable. This enabled us to assess whether the results accorded with those of Experiment 2b--namely that lexical primes associated with the compound modifier showed faster onset times. It is also important to note that, in addition to differing in terms of text vs. single word processing, this experiment differed from Experiment 2b in two other ways: First, in this experiment, both the prime and that target were typed, whereas, in Experiment 2b, only the target was typed. Second, in this experiment, the prime and target were visible from the outset of the trial, whereas, in Experiment 2b, the prime was removed from the screen before the target was shown.

To model the effect of priming of typing onset in text, a model was created using the same predictors from Experiment 2b. We also included two additional control variables. These were the average initial keystroke latency for the compound word in Experiment 2a, labelled “Baseline” in the model, and that word’s average naming latency from the English Lexicon Project (Balota et al., 2004). This variable was labelled “Mean Naming RT” in the model.

A summary of the model is shown in **Table 8**. As can be seen in this table, the pattern of priming effects corresponded to those seen for the typing of words in isolation. Lexical primes for the modifier constituent of the compound resulted in faster typing onset times (see **Figure 4**). We took the results of this experiment to indicate that lexical priming can be observed in two-sentence

TABLE 8 | Linear Mixed Effects modeling of typing onset latency of the modifier compound constituent for compound typed in two-sentence texts in Experiment 2c.

Fixed Effect	Estimate	Std. Err	df	*t*-value	*p*-value
(Intercept)	3.13	0.98	48	3.19	0.003
Prime Condition: lexical	−0.08	0.03	1956	−2.99	0.002
Prime Condition: semantic	−0.03	0.03	1945	−1.27	0.20
Constituent Primed: Head	−0.02	0.03	122	−0.76	0.45
Baseline	0.24	0.14	48	1.69	0.10
Mean Naming RT	<0.001	<0.001	48	2.60	0.01
Trial Order	0.002	<0.001	1974	−4.31	<0.001
Prime Condition: Lexical * Constituent Primed: Head	0.08	0.04	1953	2.21	0.03
Prime Condition: Semantic * Constituent Primed: Head	0.003	0.04	1946	0.10	0.92

text typing. We note, however, that text typing is, by its nature, more complex than single word typing and therefore also likely subject to more performance variability. In addition, we note that the typing onset measure in this text typing experiment is less likely to reflect ease of word recognition because, in contrast to the single word typing task reported in Experiment 2b, compounds in this experiment are presented long before they are typed.

Modeling of Within-Word Typing

To investigate whether the constituent boundary effects during within-word typing found in Experiments 2a and 2b, were also evident when compounds were typed as part of a text, a model containing the same predictors as in Experiment 2a and 2b was constructed.

As in Experiments 2a and 2b, there was a significant elevation in IKI latency at the constituent boundary. In addition, the effects of modifier frequency, head frequency, modifier positional family size, and trigram frequency accorded with the effects seen in Experiment 2a and 2b. *N* this experiment, however, we did not see an effect of head length. We thus conclude that the morphological structure of a compound affects typing latencies during compound processing in text, as they do when words are typed in isolation. As in the previous experiments, neither whole-word compound frequency nor transparency were significant predictors (i.e., there was no difference among the categories of TT, OT, and TO compounds). A summary of this model is shown in **Table 9**.

Ratings of texts collected at the end of each trial were analyzed to determine whether the lexical priming conditions contributed in any way to participants' perception of connectedness across sentences. Rating collected in Experiments 1 and 2c were combined. Several key control variables, such as response latency and prime effect were included. In order to adequately compare response latency across the two experiments, z-scores for the total response time of each trial were calculated on a participant-by-participant basis. These z-scores were used in the model to reflect individual variation in response latencies across trials. Since prime conditions were counterbalanced, individual priming effects for each participant could not be determined. Thus, mean response latencies for each compound in each condition was calculated in Experiments 1 and 2c. Priming effect coefficients were then obtained by dividing mean responses to each compound in the semantic and lexical conditions by those in the neutral condition. This produced a coefficient indicating the proportional difference in response latency across priming conditions that could be compared across all compounds and between both experiments.

Coherence ratings were converted to individual z-scores to capture the degree of individual rating variation from trial-to-trial. A linear mixed effect model was created for this dependent variable. Participants and items were included as random effects, and prime effect and response latency were included as control variables. The summary of the fixed effects for this model are presented in **Table 10**.

The estimates produced by this model indicate that participants perceived sentence pairs to be more connected when they exhibited lexical constituent overlap or lexical semantic association between lexical items spanning the sentence boundary. Compared to neutral, the highest relative ratings were observed in the lexical priming condition ($t = 5.9$, $p = <0.001$). A significant effect was also observed in the semantic condition ($t = 2.9$, $p = 0.005$). This pattern of results is shown in **Figure 5** and accords with our expectations. As we discuss again below, this accord raises the question of the directionality of causality. The data do not, in themselves, reveal whether it is the lexical co-activation that makes the text feel more coherent or whether more coherent texts enable greater lexical coactivation. At present, it seems to us that both may play a role. In this way, the accord between our text rating results and text processing results may reflect the interactive nature of language processing within texts.

GENERAL DISCUSSION

In this study, we set out to understand morphological effects in compound processing by examining production and priming in a more natural linguistic context-one that goes beyond the processing of words in isolation.

The core stimuli in the study were two-sentence texts. Each sentence in the text was composed of six words. In each text, the second word of the second sentence was a compound noun. These compound words were the focus of our study.

We constructed the two-sentence texts so that we could control the relationship between the last word of the first sentence and the second word of the second sentence (which was always a compound noun). There were three types of relationships: neutral (e.g., *path* → *waterfall*), semantic (e.g.,

TABLE 9 | Linear Mixed Effects modeling of inter-keystroke interval latency at positions relative to the morpheme boundary (position −2 is the intercept) for compound typing in sentence context in Experiment 2c.

Fixed Effect	Estimate	Std. Err	df	*t*-value	*p*-value
(Intercept)	5.74	0.11	126	53.1	<0.001
Typing Position: −1	0.03	0.01	12,500	1.85	0.06
Typing Position: Boundary	0.54	0.01	12,650	39.37	<0.001
Typing Position: +1	0.01	0.01	12,680	0.78	0.44
Typing Position: +2	0.01	0.01	12,670	0.78	0.44
Head Constituent Length	0.02	0.01	53	1.20	0.24
Trigram Frequency (log)	−0.08	0.01	6,776	−8.53	<0.001
Modifier Word Frequency	−0.03	0.02	51	−1.09	0.28
Modifier Positional Family Size	−0.003	0.002	51	−2.07	0.04
Head Word Frequency	−0.08	0.02	51	−4.12	<0.001

TABLE 10 | Summary of the linear mixed effect model of participants ratings of sentence pairs (prime condition: neutral is on the intercept).

Fixed Effect	Estimate	Std. Err	df	*t*-value	*p*-value
(Intercept)	4.22	0.10	77	44.42	<0.001
Prime Condition: Lexical	0.17	0.03	4,477	6.68	<0.001
Prime Condition: Semantic	0.06	0.03	4,471	2.49	0.01
Response Latency (Z-score)	−0.05	0.01	4,472	−3.36	<0.001
Experiment: Sentence Typing (Exp 2c)	−0.39	0.12	67	−3.25	0.002

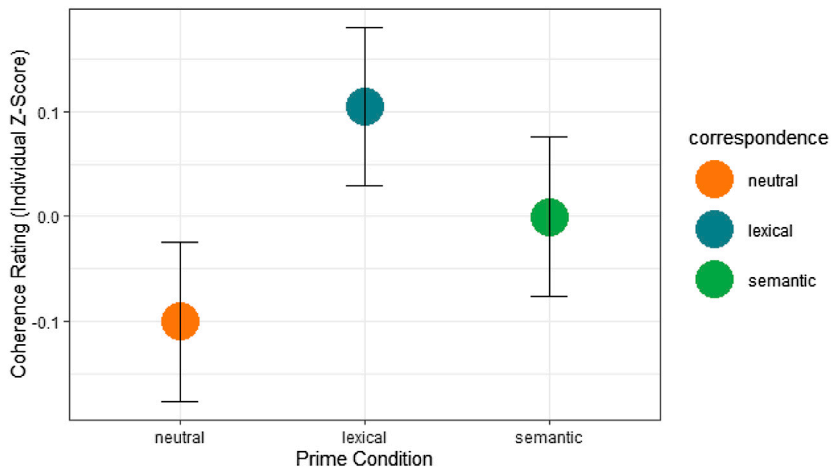


FIGURE 5 | Model estimates of sentence pair coherence rating by prime condition. Ratings are based on z-scores for each participant, indicating how much each rating deviated from that participant's mean rating.

river → *waterfall*), and lexical (e.g., *water* → *waterfall*). In Experiment 1, a maze task was employed to determine whether the speed with which the word such as *waterfall* could be chosen as a correct sentence continuation would differ depending on whether it was preceded in the text by a neutral, semantically related, or lexically related antecedent. Results indicated significant differences among all three levels, with the fastest response times associated with lexical relations. In Experiments 2a, 2b, and 2c, this finding was further investigated within a typing paradigm. In Experiment 2a, we verified that the compound words in our minimal texts showed a typing profile in which the time required to type individual medial letters within a word is greatest at the boundary between morphological constituents. In other words, in the typing of the compound word *waterfall*, it takes longer to type the letter "f" than it does to type adjacent letters. We interpreted this to constitute a signature of morphological processing – a signature that was present irrespective of the semantic transparency of the compound. In Experiment 2b, we verified that this signature remains under all conditions of priming (neutral, semantic, lexical). In addition, we found that production onset (i.e., the speed with which the first letter of the compound word was typed) was faster in the lexical condition than in either the semantic or neutral conditions (which did not differ from each other).

Having established, through Experiments 2a and 2b, the typing profiles for the core compound stimuli in isolation, we were in a position to examine whether those profiles persisted in the processing of text. This was the goal of Experiment 2c. In this experiment, participants were presented with the entire two-sentence texts and were asked to type them. We found that the morphological boundary effects seen in the typing of isolated words were replicated in the typing of words in context. We also found lexical priming effects in typing onset that were parallel to those found in single compound word typing (Experiment 2b). We thus conclude that the typing task can be used to investigate priming relations for compound words presented both in isolation and in text and that the morphological boundary signature that we observed in the typing of single compound words is also found in the typing of text. In our view, this constitutes evidence that the role of compound constituents as planning units in production is not restricted to single word processing. Indeed, our expectation is that the underlying dynamic would be in the opposite direction. It seems to us that morphological structure enables the language production system to reduce the size of the unit to be produced. Thus, morphology may play a greater role in situations in which processing demands are increased (e.g., in the free composition of text).

Turning our attention to priming, although we found constituent facilitation effects in text typing that bear similarity to the constituent priming effects found in visual priming experiments, it seems to us that the priming paradigm – so prevalent in the single-word processing literature – may have different underlying dynamics when transposed to a text processing environment. This is even more so when it is text production that is being considered. One reason for this is that the typed production task is not primarily a recognition task in the way that, for example, lexical decision is. The traditional interpretation of priming effects in a lexical decision task is that priming facilitates the "activation" of a target word. In this way, it can be considered to be a bottom-up facilitator. In our two-sentence texts, we expect that such bottom-up effects would be quite weak. On average, more than 2 s elapsed between the onset of the prime (as the last word of sentence 1) and the target (as the second word of sentence 2). However, because it is text that was being processed, it seems to us to be reasonable to posit that a function of a prime word in our minimal texts (and in a more generalized text environment) would be to set up expectations that can be described as "top-down" processing.

This leads us to a consideration of how the properties of text processing may be related to the processing of smaller linguistic units such as words and word constituents. Our first relevant finding in this domain of inquiry is that the priming patterns that we had embedded into the sentence stimuli, affected not only the lexical processing of the compounds under investigation, but also the ratings of the text as a whole. Specifically, lexical primes were associated with the highest text ratings. This was followed by semantic primes and then by neutral primes. These findings underline, in our view, the fact that although language activity can be analyzed at distinct levels (e.g., at the level of word structure, sentence structure, and text structure), real-world language processing is characterized by the integration of information and the creation of meaning.

Methodological Considerations

It seems to us that the modified maze task that we employed in Experiment 1 has great potential to unlock the interplay of lexical processing and text processing effects. This paradigm, introduced by Gallant and Libben (2020) builds upon the maze task created by Forster et al. (2009). As noted by Forster (2010), the maze task provides more reliable data than self-paced reading in cases in which there is a need to isolate processing times at specific text locations. In our study that need was critical, as we wished to use the task to investigate whether prime condition differences in Sentence 1 influenced target word processing in Sentence 2.

The use of the typing task in Experiments 2a, 2b, and 2c demonstrates the ease with which the task enables researchers to "scale up" experiments, while retaining task similarity and dependent variable comparability. Moreover, the typing task is naturally suited to studies that focus on incremental processing. Because it allows for within-word measurements, it is naturally suited to the investigation of morphological processing which is, by definition, a within-word phenomenon. The typing task also opens up new opportunities for conceptualizing the dynamics of lexical processing. For example, throughout this study, we have conceptualized morphological processing in compound production as advantageous. Yet, it seems that the signature of that advantage is a decrease in speed –perhaps best characterized as a dysfluency at particular locations. In this way, typing offers us the opportunity to relate the location of demands on computational resources to properties of the language produced and thus, potentially, to ongoing psycholinguistic processes.

Our maze task finding corresponded to the expected pattern of processing ease. The target compound word was selected with greatest speed when it was preceded by a constituent prime. This was followed by the condition in which the compound was preceded by a semantic prime. The neutral prime condition resulted in the longest maze task latencies.

CONCLUSION

This study examined three questions: 1) Are priming effects found in the processing of individual compound words also evident when the compound is part of a text and the putative prime is an antecedent word within that text? 2) Can the typing paradigm be used to examine constituent priming effects in the processing of compound words? 3) Are the elevated keystroke latencies found at the morphological boundary in the typing of single compound words also evident when the compound is part of a text?

Our interpretation of the data obtained from four experiments is that the answer to all the questions is "yes." We found corresponding constituent priming effects and morphological processing effects across experiments and, indeed, almost identical patterns in the analysis of maze task speed and the analysis of text ratings. We see this as suggesting the confluence of text effects and lexical effects (which, themselves typically bundle together form and meaning overlap). Thus, it may be advantageous to consider priming in sentence and text environments as more a matter of "interactive fit", which, by definition cannot be unidirectional.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Brock University Research Ethics Board. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors have contributed to the research design, to the development of stimuli, and to the writing of this paper.

FUNDING

This research was supported by the Social Sciences and Humanities Research Council of Canada Partnership Grant 895-2016-1008 ("Words in the World").

ACKNOWLEDGMENTS

We wish to thank Tamara Irsa-Timofeev for her stimulus development, testing and analysis that led to the study of text processing in German reported in Irsa, Dressler and Libben (2016) as well as her subsequent work on English. We also wish to thank Katharina Korecky-Kröll from the Comparative Psycholinguistics Working Group of the Department of Linguistics of University of Vienna for her support. We thank Andrea Versteeg at Brock University for her work in the development and piloting of earlier versions of English stimuli that led up to this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2021.646454/full#supplementary-material.

REFERENCES

Baayen, H., and Schreuder, R. (1999). War and peace: morphemes and full forms in a noninteractive activation parallel dual-route model. *Brain Lang.* 68, 27–32. doi:10.1006/brln.1999.2069

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., and Yap, M. J. (2004). Visual word recognition of single-syllable words. *J. Exp. Psychol. Gen.* 133, 283–316. doi:10.1037/0096-3445.133.2.283

Bauer, L. (2016). *Compounds and compounding*. Cambridge, United Kingdom: Cambridge University Press.

Beaugrande, R. D., and Dressler, W. U. (1981). *Introduction to text linguistics*. London, United Kingdom: Longman.

Clinton, V., Carlson, S. E., and Seipel, B. (2015). Linguistic markers of inference generation while reading. *J. Psycholinguist. Res.* 45 (3), 553–574. doi:10.1007/s10936-015-9360-8

Davis, C. P., Libben, G., and Segalowitz, S. J. (2019). Compounding matters: Event-related potential evidence for early semantic access to compound words. *Cognition* 184, 44–52.

Dederding, H-M. (1983). Wortbildung und Text. Zur Textfunktion (TF) von Nominalkomposita (NK). *Zeitschrift für Germanistische Linguistik* 11, 49–65. doi:10.1515/zfgl.1983.11.1.49

Dressler, W. U. (2006). "Compound types," in *The representation and processing of compound words*. Editors G. Libben and G. Jarema (Oxford: Oxford University Press), 23–44.

Ehrlich, M.-F. (1991). The processing of cohesion devices in text comprehension. *Psychol. Res.* 53, 169–174. doi:10.1007/bf01371825

Fiorentino, R., and Fund-Reznicek, E. (2009). Masked morphological priming of compound constituents. *Ment. Lexicon* 4 (2), 159–193. doi:10.1075/ml.4.2.01fio

Forster, K. I. (2010). Using a maze task to track lexical and sentence processing. *Ment. Lexicon* 5, 347–357. doi:10.1075/ml.5.3.05for

Forster, K. I., and Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *J. Exp. Psychol. Learn. Mem. Cogn.* 10, 680–698. doi:10.1037/0278-7393.10.4.680

Forster, K. I., Guerrera, C., and Elliot, L. (2009). The maze task: measuring forced incremental sentence processing time. *Behav. Res. Methods* 41, 163–171. doi:10.3758/brm.41.1.163

Gagné, C. L., and Spalding, T. L. (2014). Typing time as an index of morphological and semantic effects during English compound processing. *Lingue e linguaggio* 13 (2), 241–262. doi:10.1418/78409

Gagné, C. L., and Spalding, T. L. (2009). Constituent integration during the processing of compound words: does it involve the use of relational structures? *J. Mem. Lang.* 60 (1), 20–35. doi:10.1016/j.jml.2008.07.003

Gagné, C. L., and Spalding, T. L. (2016). Effects of morphology and semantic transparency on typing latencies in English compound and pseudocompound words. *J. Exp. Psychol. Learn. Mem. Cogn.* 42, 1489–1495. doi:10.1037/xlm0000258

Gallant, J., and Libben, G. (2020). Can the maze task be even more amazing? Adapting the maze task to advance psycholinguistic experimentation. *The Mental Lexicon* 15 (2), 366–383.

Giraudo, H., and Grainger, J. (2001). Priming complex words: evidence for supralexical representation of morphology. *Psychon. Bull. Rev.* 8, 127–131. doi:10.3758/bf03196148

Goral, M., Libben, G., Obler, L. K., Jarema, G., and Ohayon, K. (2008). Lexical attrition in younger and older bilingual adults. *Clin. Linguist. Phon.* 22 (7), 509–522. doi:10.1080/02699200801912237

Graesser, A. C., MillisMillis, K. K., and Zwaan, R. A. (1997). Discourse comprehension. *Annu. Rev. Psychol.* 48, 163–189. doi:10.1146/annurev.psych.48.1.163

Graesser, A. C., Singer, M., and Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychol. Rev.* 101 (3), 371. doi:10.1037/0033-295x.101.3.371

Huang, J., Tsang, Y-K., Xiao, W., and Wang, S. (2020). Morphosemantic activation of opaque Chinese words in sentence comprehension. *PLoS One* 15 (8), e0236697. doi:10.1371/journal.pone.0236697

Irsa, T., Dressler, W. U., and Libben, G. (2016). "Competition and cooperation between top-down and bottom-up processing of compounds," in Paper presented at the 17th International Morphology Meeting, Vienna, Austria, February 18–February 21, 2016.

Ji, H., Gagné, C. L., and Spalding, T. L. (2011). Benefits and costs of lexical decomposition and semantic integration during the processing of transparent and opaque English compounds. *J. Mem. Lang.* 65 (4), 406–430. doi:10.1016/j.jml.2011.07.003

Kintsch, W. (1998). *Comprehension: a paradigm for cognition*. Cambridge, UK: Cambridge University Press.

Kintsch, W., and van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychol. Rev.* 85, 363–394. doi:10.1037/0033-295x.85.5.363

Kuperman, V., Schreuder, R., Bertram, R., and Baayen, R. H. (2009). Reading polymorphemic Dutch compounds: toward a multiple route model of lexical processing. *J. Exp. Psychol. Hum. Perception Perform.* 35, 876–895. doi:10.1037/a0013484

Lenth, R. (2020). emmeans: estimated marginal means, aka least-squares means. package version 1.5.0NAvailable at: https://CRAN.R-project.org/package=emmeans.

Libben, G., Curtiss, K., and Weber, S. (2014). Psychocentricity and participant profiles: implications for lexical processing among multilinguals. *Front. Psychol.* 5, 557. doi:10.3389/fpsyg.2014.00557

Libben, G., Gagné, C., and Dressler, W. U. (2020). "The representation and processing of compounds words," in *Word knowledge and word usage: a cross-disciplinary guide to the mental Lexicon*. Editors V. Pirreli, I. Plag, and W. U. Dressler (Berlin, Germany: De Gruyter Mouton).

Libben, G., Jarema, G., Derwing, B., Riccardi, A., and Perlak, D. (2016). Seeking the -ationalinderivationalmorphology. *Aphasiology* 30, 1304–1324. doi:10.1080/02687038.2016.1165179

Libben, G. (2006). "Why study compounds? An overview of the issues," in *The representation and processing of compound words*. Editors G. Libben and G. Jarema (Oxford: Oxford University Press), 1–21.

Libben, G. (2014). The nature of compounds: a psychocentric perspective. *Cogn. Neuropsychol.* 31, 8–25. doi:10.1080/02643294.2013.874994

Libben, G., and Weber, S. (2014). "Semantic transparency, compounding, and the nature of independent variables," in *Morphology and meaning*. Editors F. Rainer, U. Wolfgang, F. Gardani, W. U. Dressler, and H. C. Luschützky (Amsterdam, Netherlands: John Benjamins).

Pollatsek, A., Hyönä, J., and Bertram, R. (2000). The role of morphological constituents in reading Finnish compound words. *J. Exp. Psychol. Hum. Perception Perform.* 26 (2), 820. doi:10.1037/0096-1523.26.2.820

Marelli, M., Crepaldi, D., and Luzzatti, C. (2009). Head position and the mental representation of nominal compounds. *Ment. Lexicon* 4 (3), 430–454. doi:10.1075/ml.4.3.05mar

Monahan, P. J., Fiorentino, R., and Poeppel, D. (2008). Masked repetition priming using magnetoencephalography. *Brain Lang.* 106 (1), 65–71. doi:10.1016/j.bandl.2008.02.002

Mousikou, P., and Schroeder, S. (2019). Morphological processing in single-word and sentence reading. *J. Exp. Psychol. Learn. Mem. Cogn.* 45, 881–903. doi:10.1037/xlm0000619

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., et al. (2019). PsychoPy2: experiments in behavior made easy. *Behav. Res. Methods* 51 (1), 195–203. doi:10.3758/s13428-018-01193-y

Sahel, S., Nottbusch, G., Grimm, A., and Weingarten, R. (2008). Written production of German compounds: effects of lexical frequency and semantic transparency. *Written Lang. Literacy* 11 (2), 211–227. doi:10.1075/wll.11.2.06sah

Sandra, D. (2019). *Morphological units: a theoretical and psycholinguistic perspective. Oxford encyclopedia of linguistics*. Oxford, UK: Oxford University Press.

Sandra, D. (1990). On the representation and processing of compound words: automatic access to constituent morphemes does not occur. *Q. J. Exp. Psychol.* 42 (3), 529–567. doi:10.1080/14640749008401236

Smith, V., Barratt, D., and Zlatev, J. (2014). Unpacking noun-noun compounds: interpreting novel and conventional foodnames in isolation and on food labels. *Cogn. Linguist.* 25 (1), 99–147. doi:10.1515/cog-2013-0032

Taft, M., and Forster, K. I. (1976). Lexical storage and retrieval of polymorphemic and polysyllabic words. *J. Verbal Learning Verbal Behav.* 15 (16), 607–620.

Verhoeven, L., and Perfetti, Ch.. (2008). Advances in text comprehension: model, process and development. *Appl. Cogn. Psychol.* 22,.293–301. doi:10.1002/acp.1417

Weingarten, R., Nottbusch, G., and Will, U. (2004). Morphemes, syllables, and graphemes in written word production. *Trends Linguist. Studies Monographs* 157, 529–572. doi:10.1515/9783110894028.529

Wurm, L. H., and Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *J. Mem. Lang.* 72, 37–48. doi:10.1016/j.jml.2013.12.003

Zwitserlood, P. (1994). The role of semantic transparency in the processing and representation of Dutch compounds. *Lang. Cogn. Process.* 9 (3), 341–368. doi:10.1080/01690969408402123

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Frontiers in Communication

Investigates the power of communication across culture and society

A cross-disciplinary journal that advances our understanding of the global communication revolution and its relevance across social, economic and cultural spheres.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact