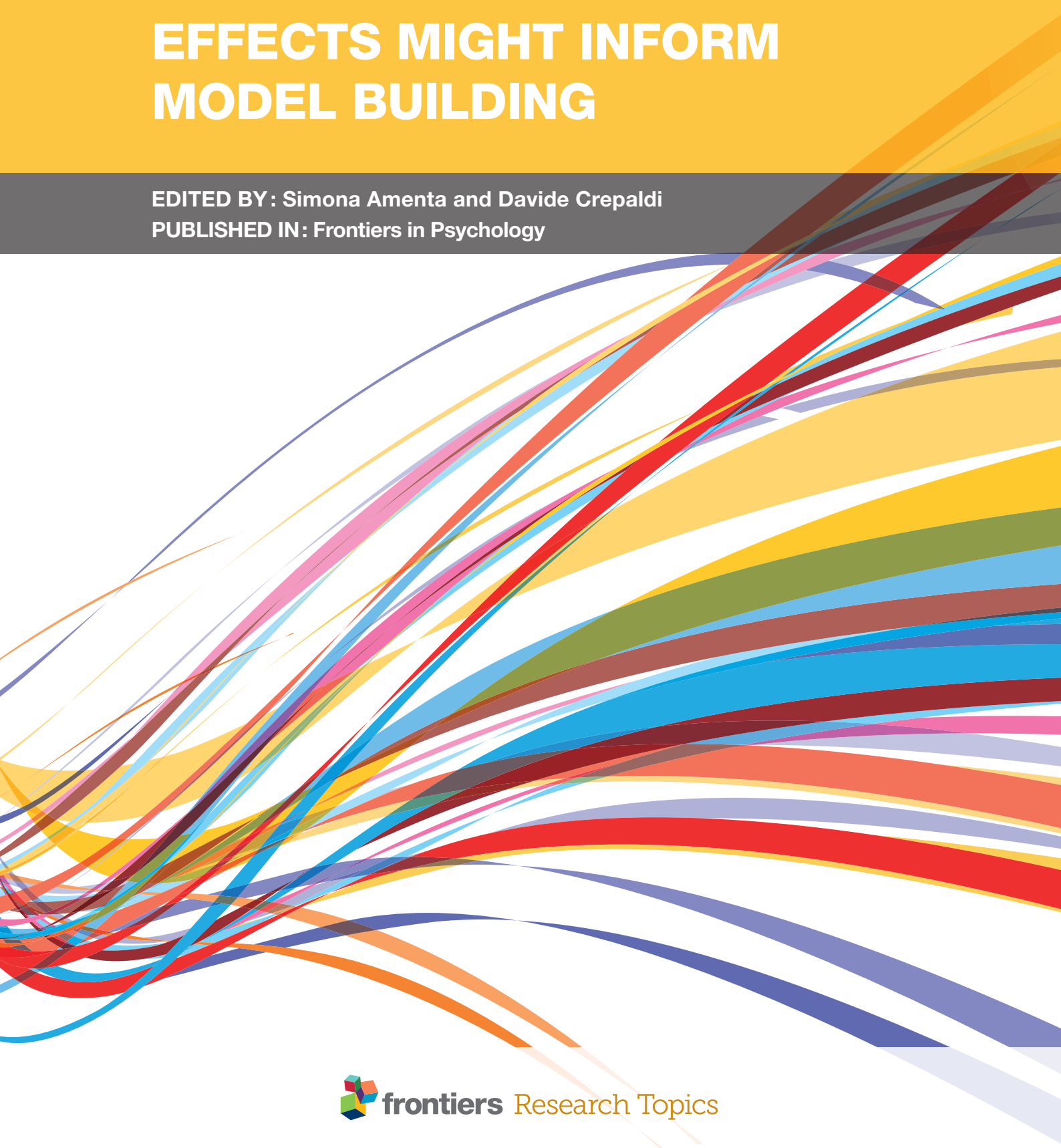
THE VARIABLE MIND? HOW APPARENTLY INCONSISTENT EFFECTS MIGHT INFORM MODEL BUILDING

EDITED BY: Simona Amenta and Davide Crepaldi
PUBLISHED IN: Frontiers in Psychology



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88919-859-7
DOI 10.3389/978-2-88919-859-7

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: **researchtopics@frontiersin.org**

THE VARIABLE MIND? HOW APPARENTLY INCONSISTENT EFFECTS MIGHT INFORM MODEL BUILDING

Topic Editors:
Simona Amenta, University of Trento, Italy
Davide Crepaldi, International School for Advanced Studies & Milan Center for Neuroscience, Italy

Model building is typically based on the identification of a set of established facts in any given field of research, insofar as the model is then evaluated on how well it accounts for these facts. Psychology – and specifically visual word identification and reading – is no exception in this sense (e.g., Amenta & Crepaldi, 2012; Coltheart et al., 2001; Grainger & Jacobs, 1996).

What counts as an established fact, however, was never discussed in great detail. It was typically considered, for example, that experimental effects need to replicate across, e.g., individuals, experimental settings, and languages if they are to be believed. The emphasis was on consistency, perhaps under a tacit assumption that the universal principles lying behind our cognitive structures determine our behaviour for the most part (or at least for that part that is relevant for model building).

There are signs that a different approach is growing up in reading research. On a theoretical ground, Dennis Norris' Bayesian reader (2006, 2009) has advanced the idea that models can dispense of static forms of representation (i.e., fixed architectures), and process information in a way that is dynamically constrained by context-specific requirements. Ram Frost (2012) has focused on language-specific constraints in the development of general theories of reading. On an empirical ground, the most notable recent advance in visual word identification concern the demonstration that some previously established (in the classic sense) effects depend heavily on language (Velan and Frost, 2011), task (e.g., Duñabeitia et al., 2011; Marelli et al., 2013; Kinoshita and Norris, 2009), or even individual differences (Andrews & Lo, 2012, 2013). Variability has become an intrinsic and informative aspect of cognitive processing, rather than a sign of experimental weakness.

This Research Topic aims at moving forward in this new direction by providing an outlet for experimental and theoretical papers that: (i) explore more in depth the theoretical basis for considering variability as an intrinsic property of the human cognitive system; (ii) highlight

new context-dependent experimental effects, in a way that is informative on the dynamics of the underlying cognitive processing; (iii) shed new light on known context-dependent experimental effects, again in a way that enhances their theoretical informativeness.

Citation: Amenta, S., Crepaldi, D., eds. (2016). The Variable Mind? How Apparently Inconsistent Effects Might Inform Model Building. Lausanne: Frontiers Media. doi: 10.3389/ 978-2-88919-859-7

Table of Contents

05 ***Editorial: The Variable Mind? How Apparently Inconsistent Effects Might Inform Model Building***
Simona Amenta and Davide Crepaldi

07 ***Framing effects reveal discrete lexical-semantic and sublexical procedures in reading: an fMRI study***
Laura Danelli, Marco Marelli, Manuela Berlingeri, Marco Tettamanti, Maurizio Sberna, Eraldo Paulesu and Claudio Luzzatti

25 ***Item parameters dissociate between expectation formats: a regression analysis of time-frequency decomposed EEG data***
Irene F. Monsalve, Alejandro Pérez and Nicola Molinaro

37 ***How language affects children's use of derivational morphology in visual word and pseudoword processing: evidence from a cross-language study***
Séverine Casalis, Pauline Quémart and Lynne G. Duncan

47 ***Does the mean adequately represent reading performance? Evidence from a cross-linguistic study***
Chiara V. Marinelli, Joanna K. Horne, Sarah P. McGeown, Pierluigi Zoccolotti and Marialuisa Martelli

63 ***List context effects in languages with opaque and transparent orthographies: a challenge for models of reading***
Daniela Traficante and Cristina Burani

76 ***An ERP study of effects of regularity and consistency in delayed naming and lexicality judgment in a logographic writing system***
Yen Na Yum, Sam-Po Law, I-Fan Su, Kai-Yan Dustin Lau and Kwan Nok Mo

88 ***Measuring inconsistencies can lead you forward: Imageability and the x-ception theory***
Sara Dellantonio, Claudio Mulatti, Luigi Pastore and Remo Job

97 ***Hierarchical clustering analysis of reading aloud data: a new technique for evaluating the performance of computational models***
Serje Robidoux and Stephen C. Pritchard

104 ***Relative clause reading in hearing impairment: different profiles of syntactic impairment***
Ronit Szterman and Naama Friedmann

120 ***Colors, colored overlays, and reading skills***
Arcangelo Uccula, Mauro Enna and Claudio Mulatti

124 ***Is there a bilingual advantage in the ANT task? Evidence from children***
Eneko Antón, Jon A. Duñabeitia, Adelina Estévez, Juan A. Hernández, Alejandro Castillo, Luis J. Fuentes, Douglas J. Davidson and Manuel Carreiras

frontiers in Psychology

EDITORIAL
published: 22 February 2016
doi: 10.3389/fpsyg.2016.00185

Editorial: The Variable Mind? How Apparently Inconsistent Effects Might Inform Model Building

Simona Amenta[1]* *and Davide Crepaldi*[2,3]

[1] Center for Mind/Brain Sciences, University of Trento, Rovereto, Italy, [2] International School for Advanced Studies, Trieste, Italy, [3] Milan Center for Neuroscience, Milan, Italy

Keywords: experimental variability, individual differences, reading, bilingualism, dyslexia

The Editorial on the Research Topic

The Variable Mind? How Apparently Inconsistent Effects Might Inform Model Building

Human behavior is very difficult to predict precisely, as even the exact same cognitive system may respond differently to very similar input. That is, the study of experimental psychology and neuroscience requires dealing with a huge amount of variability. Our response to this state of affairs, as a field, has been dominated by the (rather tacit) assumption that variability means noise, and thus it is something we need to (i) ignore theoretically; and (ii) fight against experimentally, searching for stable effects. Although it remains obvious that part of the variability we see in our experiments is indeed noise, a different approach emerged recently, based on the assumption that the cognitive system is guided by dynamic and flexible architectures that adapt quickly to different contexts. Thus, how psychological effects emerge and disappear in different, e.g., people, contexts, languages, brings light into the features of the cognitive system itself (e.g., Norris, 2006; Andrews and Lo, 2013). Variability is in focus as an intrinsic aspect of cognitive processing, rather than a sign of experimental weakness; and the experimental and theoretical enterprise is directed toward the validation of consistently variable facts. The present E-book is a collection of experimental and theoretical work that moves in this direction, focusing on how variability may inform theoretical advance.

The focus on variability has been interpreted in terms of context effects by Danelli et al. and Monsalve et al. The former group conducted an fMRI experiment where similar sets of regular words were presented to participants together with either nonwords or irregular words, in an attempt to enhance grapheme–to–phoneme or lexical–semantic reading respectively, in the context of dual–route models of reading (e.g., Coltheart et al., 2001). Brain activations were partially different in the two contexts, thus leading Danelli et al. to claim association between different neural circuits and either sub–lexical or lexical reading. Monsalve et al. compared ERPs elicited by the same words when embedded in: (i) multiword expressions (e.g., "kick the bucket"); (ii) highly predictable, but non–fixed compositional structures (e.g., "the opposite of black is white"); or (iii) non–constraining contexts (e.g., "Phil asked Mary to bring her ring"). The same exact set of words brought about different neural responses in different contexts, in this case teasing apart lexical identification and word prediction.

Variability across languages is another hot issue in this Research Topic. Focusing on English and French, Casalis et al. assessed the role of morphemes in the reading performance of a group of children. In both languages, the presence of derivational morphemes facilitates word recognition (e.g., "postal" better than "turnip"), and hinders nonword rejection (e.g., "pondal" worse than "curlip"). However, the same factor affects latencies in the two languages in different ways, possibly due to the different derivational structures of English and French.

OPEN ACCESS

Edited and reviewed by:
Manuel Carreiras, Basque Center on Cognition, Brain and Language, Spain

***Correspondence:**
Simona Amenta
simona.amenta@gmail.com

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

Received: *28 January 2016*
Accepted: *31 January 2016*
Published: *22 February 2016*

Citation:
Amenta S and Crepaldi D (2016) Editorial: The Variable Mind? How Apparently Inconsistent Effects Might Inform Model Building. Front. Psychol. 7:185. doi: 10.3389/fpsyg.2016.00185

Differences between language orthographies are explored in the study by Marinelli et al., who compare reading performance of English and Italian young adults in a series of three experiments. By means of an ex–Gaussian distribution analysis, the authors unveiled more diversity among English readers in terms of response time variability and the amount of slow responses; but not in terms of mean, which is what goes under the microscope more often. Traficante and Burani contributed a review that focuses on how different orthographic systems (particularly in terms of letter–to–sound mapping consistency) shape different ways in which the lexical and sub–lexical reading routes are (de)emphasized according to context. Yum et al. built on the special features of Chinese to tease apart regularity (how much the pronunciation of a word follows letter–to–sound conversion rules) and consistency (how much the pronunciation of a word is consistent with that of words with similar orthography). In a series of ERP studies, the authors find indeed different timings and different directions for the electrophysiological effects of the two constructs.

Dellantonio et al. focused instead on differences between types of words within the same language. Building on the MRC database (Coltheart, 1981), they were able to identify groups of words where the classic correlation between imageability and concreteness ratings doesn't hold, thus helping clarifying the difference between the two constructs.

Inter–individual variability is the target in Robidoux and Pritchard, who compared the responses predicted by the DRC (Coltheart et al., 2001) and the CDP++ (Perry et al., 2010) models of reading to those of a group of human subjects. By means of hierarchical clustering, they individuated groups of subjects that differ in the pronunciation of specific consonant clusters. Based on this finding, Robidoux and Pritchard compared DRC and CDP++ for their ability to model different, but internally consistent, reading profiles, setting a new and interesting way to address the long–lasting issue of adjudicating between different reading models.

Individual differences obviously go well beyond different mappings between sounds and graphemes in fully functional adults. Szterman and Friedmann, for example, analyzed the difficulties that children with impaired hearing show with Wh-movement sentences, highlighting differences in syntactic processing mainly related to the use of a hearing device within the first year of life. Uccula et al. focused instead on the Meares–Irlen syndrome, a condition whereby readers experience eyestrain and/or visual distortions, and reading improves quite dramatically through the use of colored overlays applied above written text.

Finally, the ability to speak more than one language has been quite consistently linked to the ability to outperform monolinguals in a variety of tasks (e.g., Bialystok, 2001). However, by adopting new Bayesian analyses on a large sample, Antón et al. were able to question this connection showing that, in a series of tasks, performance of bilinguals and monolinguals is statistically undistinguishable.

Overall, we are confident that this Research Topic provides solid examples of how consistent variability can inform psychological theory. Contributions cover diverse populations, as well as diverse techniques, which proves that language psychology is indeed widening its toolbox by including the study of how phenomena vary across tasks, contexts, languages, words and people. We hope that this move, still in its early phase, will consolidate and provide more and more insight into the beauty of the human cognitive system.

AUTHOR CONTRIBUTIONS

SA and DC equally shared the task of editing the content of this Research Topic, and the authorship of this Editorial.

FUNDING

This work was funded by a "FIRB-Futuro in Ricerca" grant awarded to DC by the Italian Ministry of Education, University and Research (RBFR085K98).

ACKNOWLEDGMENTS

This work was partially carried out when the Editors were affiliated to the Department of Psychology at the University of Milano-Bicocca. SA and DC wish to thank the 61 scholars that contributed as authors, reviewers, or editors to this work granting the completion of the Research Topic.

REFERENCES

Andrews, S., and Lo, S. (2013). Is morphological priming stronger for transparent than opaque words? It depends on individual differences in spelling and vocabulary. *J. Mem. Lang.* 68, 279–296. doi: 10.1016/j.jml.2012.12.001

Bialystok, E. (2001). *Bilingualism in Development: Language, Literacy, and Cognition*. New York, NY: Cambridge University Press.

Coltheart, M. (1981). The MRC psycholinguistic database. *Q. J. Exp. Psychol.* 33, 497–505.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., and Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychol. Rev.* 108, 204–256. doi: 10.1037/0033-295X.108.1.204

Norris, D. (2006). The Bayesian reader: explaining word recognition as an optimal Bayesian decision process. *Psychol. Rev.* 113, 327–357. doi: 10.1037/0033-295X.113.2.327

Perry, C., Ziegler, J. C., and Zorzi, M. (2010). Beyond single syllables: Large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. *Cogn. Psychol.* 61, 106–151. doi: 10.1016/j.cogpsych.2010.04.001

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: 23 September 2015
doi: 10.3389/fpsyg.2015.01328

Framing effects reveal discrete lexical-semantic and sublexical procedures in reading: an fMRI study

Laura Danelli[1,2], Marco Marelli[1,3], Manuela Berlingeri[1,2], Marco Tettamanti[4], Maurizio Sberna[5], Eraldo Paulesu[1,2,6] and Claudio Luzzatti[1,2]*

[1] Psychology Department, University of Milan-Bicocca, Milan, Italy, [2] NeuroMI -Milan Center for Neuroscience, Milan, Italy, [3] Center for Mind/Brain Sciences, University of Trento, Rovereto, Italy, [4] Division of Neuroscience and Department of Nuclear Medicine, San Raffaele Scientific Institute, Milan, Italy, [5] Neuroradiology Department, Niguarda Ca' Granda Hospital, Milan, Italy, [6] fMRI Unit, IRCCS Galeazzi, Milan, Italy

OPEN ACCESS

Edited by:
Manuel Carreiras, Basque Center on Cognition, Brain and Language, Spain

Reviewed by:
Jeremy Purcell, Georgetown University, USA
Remo Job, University of Trento, Italy

***Correspondence:**
Laura Danelli, Dipartimento di Psicologia, Università degli Studi di Milan-Bicocca, Pizza dell'Ateneo Nuovo, 1, 20126 Milan, Italy
laura.danelli@unimib.it

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

Received: *15 April 2014*
Accepted: *18 August 2015*
Published: *23 September 2015*

Citation:
Danelli L, Marelli M, Berlingeri M, Tettamanti M, Sberna M, Paulesu E and Luzzatti C (2015) Framing effects reveal discrete lexical-semantic and sublexical procedures in reading: an fMRI study. Front. Psychol. 6:1328. doi: 10.3389/fpsyg.2015.01328

According to the dual-route model, a printed string of letters can be processed by either a grapheme-to-phoneme conversion (GPC) route or a lexical-semantic route. Although meta-analyses of the imaging literature support the existence of distinct but interacting reading procedures, individual neuroimaging studies that explored neural correlates of reading yielded inconclusive results. We used a list-manipulation paradigm to provide a fresh empirical look at this issue and to isolate specific areas that underlie the two reading procedures. In a lexical condition, we embedded disyllabic Italian words (target stimuli) in lists of either loanwords or trisyllabic Italian words with unpredictable stress position. In a GPC condition, similar target stimuli were included within lists of pseudowords. The procedure was designed to induce participants to emphasize either the lexical-semantic or the GPC reading procedure, while controlling for possible linguistic confounds and keeping the reading task requirements stable across the two conditions. Thirty-three adults participated in the behavioral study, and 20 further adult participants were included in the fMRI study. At the behavioral level, we found sizeable effects of the framing manipulations that included slower voice onset times for stimuli in the pseudoword frames. At the functional anatomical level, the occipital and temporal regions, and the intraparietal sulcus were specifically activated when subjects were reading target words in a lexical frame. The inferior parietal and anterior fusiform cortex were specifically activated in the GPC condition. These patterns of activation represented a valid classifying model of fMRI images associated with target reading in both frames in the multi-voxel pattern analyses. Further activations were shared by the two procedures in the occipital and inferior parietal areas, in the premotor cortex, in the frontal regions and the left supplementary motor area. These regions are most likely involved in either early input or late output processes.

Keywords: reading, fMRI, list-manipulation paradigm, dual-route model, lexical-semantic procedure, sublexical procedure, multi-voxel pattern analysis (MVPA)

Introduction

In the study of reading, dual-route models (Coltheart et al., 1980, 2001) have been very influential in both experimental psychology and neuropsychology. These models assume that a string of letters can be processed by two procedures. One procedure is based on a GPC route to generate individual sounds and the assembled phonology represented by the whole string. The other procedure is based on a lexical, or a lexical and semantic, route that is initially activated in the form of an abstract word representation in the *orthographic input lexicon*. This would then activate the corresponding conceptual representation and the phonological word form. These two processing routes have been assumed to run in parallel, and in principle, any orthographic input representation triggers the activation of both streams. However, this does not necessarily mean that all stimuli can be read correctly along both routes. Indeed, the involvement of either route in processing specific stimuli leads to erroneous (or even impossible) outcomes. Pseudowords (e.g., *sploice*) cannot be read via the lexical route because they lack a lexical representation, and irregular words (e.g., *yacht*) are doomed to incorrect readings (regularizations) when they are processed with a GPC procedure (*yacht* read as /jɒt/). The two processing-routes hypothesis (the dual-route model) has been studied extensively in cognitive neuropsychology. Patients who are impaired in pseudoword reading and whose performance on words was flawless (phonological dyslexia; Beauvois and Derousné, 1979; Shallice and Warrington, 1980; Coltheart, 1996), and patients who correctly read pseudowords and regular words but fail when trying to read irregular words (surface dyslexia; Marshall and Newcombe, 1973) represent a double dissociation in support of the dual-route hypothesis. The former type of dyslexia can be explained as a consequence of specific damage to the GPC procedure, and the latter is interpreted in terms of an impairment of the lexical route[1]. The success of the dual-route approach in explaining neuropsychological impairment has made it a reference model in the field and an important theoretical framework for clinical assessment.

Anatomical investigation in patients with specific forms of dyslexia suggests that the cognitive procedures that are involved in either processing route could be associated with two anatomically segregated processing streams for reading abilities. Poor reading of pseudowords is usually associated with temporo-parietal and left frontal lesions (e.g., Friedman and Kohn, 1990; Friedman, 1996; Patterson et al., 1996; Fiez et al., 2006; Sato et al., 2008; Rapcsak et al., 2009), and an impairment in reading irregular words is often observed with left anterolateral temporal lobe damage (e.g., Patterson and Behrmann, 1997; Wilson et al., 2009; see also Ripamonti et al., 2014 for a voxel-based symptom mapping analysis of 59 dyslexic patients). However, the concept of independent and segregated networks that are associated with each reading route is not unequivocally accepted in the literature for both empirical and theoretical reasons. From an empirical point of view, neuropsychological results are not conclusive because most phonological and surface dyslexic patients suffer from extensive and heterogeneous lesions that make it difficult to establish well-localized functional anatomical correlations. Behavioral deficits can also occur due to either a lesion of a specific brain region or an anatomical disconnection between cerebral regions.

The dual-route model is only one of the theoretical frameworks that have been proposed in the literature to explain reading processes. The most successful alternative to the dual-route model, the connectionist model (or *triangle model*; Seidenberg and McClelland, 1989; Plaut et al., 1996; Seidenberg, 2005) suggests a strongest emphasis on the orthography-to-semantics-to-phonology pathway for irregularly spelled words and on orthography-to-phonology processes for pseudowords (for a functional anatomical demonstration see Mechelli et al., 2005). Crucially, this model does not postulate a separate, lexical non-semantic route for reading.

Functional Imaging Contributions to the Identification of Specific Pathways for Reading: The Impregnable Fortress of the Dual-route Pathway

There are several reasons why the imaging literature has failed to provide convincing evidence of dissociable neural systems for the sublexical and lexical routes (see Cattinelli et al., 2013; Taylor et al., 2013). Many strategies have been adopted. One experimental strategy has been to manipulate task demands rather than stimuli (Rumsey et al., 1997; Cappa et al., 1998; Mummery et al., 1998; Booth et al., 2002). The assumption made in these studies is that very similar items may be processed differently by varying the specific task demands. A classical implementation of this rationale has been the adoption of semantic as opposed to phonological judgment tasks for the same stimuli. The results of these two approaches would provide information about the areas that are involved in the lexical route or in the GPC route, respectively (Price et al., 1997; Rumsey et al., 1997; Mummery et al., 1998; Booth et al., 2002). This approach is complicated by the difficulty of controlling for the activation of semantic representations. A further problem in the task-demand manipulation approach is that certain cognitive tasks, such as phonological or semantic awareness tasks, tap into high-level cognitive layers that are associated with the decision-making processes and the selection of relevant information that suggests which cognitive judgments are to be made. It is plausible that this type of manipulation will strongly affect neural activation as well (see Table 6 in Cattinelli et al., 2013).

Another popular approach has been to use route-specific sets of stimuli. English orthography is an ideal test-bed because of its many orthographic irregularities. It has been assumed that *route-specific* sets of items would activate only those areas that are associated with a specific procedure. For example, pseudowords would emphasize the *GPC areas*, and irregular words would emphasize areas that are involved in the lexical procedure (see illustrative examples and reviews in Fiez and Petersen, 1998; Hagoort et al., 1999; Paulesu et al., 2000; Mechelli et al., 2003; Ino et al., 2009; Levy et al., 2009; Price, 2012; Cattinelli et al., 2013).

[1]The frequent co-occurrence of more general phonological deficits in phonological dyslexia or of semantic deficits in surface dyslexia has fuelled the debate about whether specific sublexical or lexical reading procedures actually exist (see Patterson and Ralph, 1999 for a review).

Data derived from this approach are in some cases contradictory. For example, several fMRI studies reported stronger activation of the left occipito-temporal junction (Paulesu et al., 2000; Xu et al., 2001) and of the left inferior temporal cortex (Fiez et al., 1999; Paulesu et al., 2000) during pseudoword reading compared with word reading, which would suggest involvement of these areas in sublexical processing or the contribution of larger-grained representations to pseudoword reading (see Cattinelli et al., 2013). On the contrary, some studies reported a stronger activation of these same areas when reading words rather than pseudowords (Cappa et al., 1998; Hagoort et al., 1999), which would suggest involvement of these regions in written word processing.

Furthermore, this approach is prone to possible confounds (e.g., familiarity with the orthographic string, the role of variables such as word frequency, or imageability), and it strongly depends on the assumption that brain regions that are specifically involved in a given procedure (e.g., a GPC-specific region) would be functionally silent when reading stimuli that are preferentially processed by the alternative procedure. There is evidence, however, that this might not be the case: for example, the assumption that irregular > regular words would activate the lexical route and regular > irregular words would activate the GPC route is not valid, as both types of stimuli actually activate both routes (although only one is functionally relevant, see Coltheart et al., 2001; Taylor et al., 2013).

Recently, the PET/fMRI literature has been reviewed in two meta-analyses (Cattinelli et al., 2013; Taylor et al., 2013) to address this issue. Both Cattinelli et al. (2013) and Taylor et al. (2013) found evidence for greater activity during pseudoword than word reading (which should reveal activity in brain regions that are involved in spelling-to-sound conversion) in the bilateral parietal cortex and the left posterior occipito-temporal cortex. Moreover, both studies found evidence for greater activity for word than pseudoword reading (which should reveal activity in brain regions that are involved in lexical/semantic processing) in the left angular gyrus, left anterior fusiform gyrus, and left middle temporal gyrus.

Rationale and Aim of the Present Study

The aim of this study was to challenge the *dual-route anatomical fortress*. We capitalized on previous evidence that suggests that it is possible to influence sequential single-word reading strategies by manipulating the item lists either by employing separate lists for different item types or by mixing different types of stimuli, e.g., pseudowords and words, within the same list (Baluch and Besner, 1991; Monsell et al., 1992; Tabossi and Laghi, 1992; Lupker et al., 1997; Zevin and Balota, 2000; Decker et al., 2003; Reynolds and Besner, 2005; Kinoshita and Lupker, 2007; Kang et al., 2009; Paizi et al., 2010; see Traficante and Burani, 2014 for a review). The data seem to support the assumption that manipulation of an experimental list may induce preferential recruitment of either lexical or sublexical strategies: a reading task where regular words are mixed with irregular words might lead to intensification of the lexical process, while a condition where regular words are mixed with pseudowords might emphasize the sublexical route (the route emphasis hypothesis; Monsell et al., 1992; Reynolds and Besner, 2005). However, some authors proposed an alternative interpretation of these effects, which suggests that the onset of the verbal response in pure and mixed lists could be modulated by the specific demand that is imposed by the item to be pronounced (e.g., its phonological and articulatory aspects, and whether it is more or less frequent). According to this hypothesis, reading pace would be determined by item "difficulty" and would reflect the participant's attempt to strike a balance between reading accuracy and reading speed (the time-criterion hypothesis; Lupker et al., 1997; Kinoshita and Lupker, 2007; Kang et al., 2009).

Interestingly, although they originate from the same behavioral evidence, these two interpretations result in opposite neurofunctional predictions. The route emphasis hypothesis predicts that the adoption of list manipulation would result in the recruitment of different neural patterns in response to different reading procedures, while the time-criterion hypothesis predicts that the adoption of different reading items, even if they elicit a same reading procedure, would be associated with different neural activations. In particular, according to the time-criterion hypothesis, we should be able to detect between-items differences in those brain regions that are typically associated with task demand (Bedny et al., 2007; Berlingeri et al., 2008). Item demand may depend on early visual-orthographic features, on phonological-articulatory complexity, on psycholinguistic aspects or on a combination of these different levels.

In light of these considerations, our approach was to embed disyllabic real words (target words) in a frame of non-target stimuli, which would lead participants to place more emphasis on either the lexical or the sublexical reading procedure. Irregular words[2] would emphasize the lexical procedure, and pseudowords would emphasize the GPC reading procedure. We used two classes of irregular lists (trisyllabic words with unpredictable stress position, loan words that are largely employed in the Italian language, e.g., "computer") to allow us to generalize our findings beyond one single class of stimuli. Moreover the use of two experimental conditions, one for loanwords and the other for trisyllabic words, allowed us not only to elicit the lexical-semantic reading strategy (or the GPC reading procedure with the pseudoword frames), but also to address the time-criterion hypothesis. Indeed, even if our fMRI manipulation paradigm was designed to specifically address the route-emphasis hypothesis and to disentangle the neurofunctional correlates of the lexical and sublexical routes while controlling all possible lexical and experimental confounds, the concept of item demands allowed us to assess the time-criterion hypothesis by looking for a possible interaction between experimental conditions and lexicality in the fMRI data.

In other words, if a gradient of difficulty between different filler lists would actually occur (i.e., pseudowords with CV structure easier to read than pseudowords with complex consonant clusters, and words with CV structure easier to

[2]There are no irregularities of orthography for reading in Italian. The only reading ambiguity emerges when one has to retrieve the proper stress position for words of more than two syllables. The stress position for such multisyllabic words can be retrieved only by lexical identification and not by means of orthographic cues.

read than loanwords) and this difficulty effect would influence the reading speed of target words, as assumed by the time-criterion hypothesis, an interaction effect between session and lexicality should emerge at anatomo-functional level. Whereas, if no interaction effects would emerge from fMRI data, we may infer that the list manipulation actually induces different reading procedures, thus supporting the route-emphasis hypothesis.

Finally, it is worthy to note that the target stimuli for the analyses of the hemodynamic responses were always disyllabic words. These were accurately matched for a number of psycholinguistic properties, such as word frequency, phonological complexity, orthographic neighborhood size, imageability, and beginning phonemes. With this new experimental paradigm, we tried to find evidence for process-specific areas and for brain regions that are shared by the two reading procedures, and simultaneously we tried to control for possible psycholinguistic and task-related confounds.

Materials and Methods

Behavioral Study

Participants

Thirty-three healthy young adult participants (17 M/16 F; mean age = 28.6 ± 4.4 years) took part in the study. The participants had no history of neurological and psychiatric disorders. They all had normal cognitive development, normal or corrected-to-normal vision, and normal language and reading skills.

Stimuli

Regular disyllabic Italian words (target stimuli) were embedded in a frame of either irregular words or pseudowords (fillers). There were four lists, and each comprised 20 disyllabic Italian words. All experimental words were nouns with a consonant-vowel (CV) structure. The four lists were matched for word frequency, orthographic neighborhood size, imageability, and beginning phonemes. Word frequency and orthographic neighborhood size measures were obtained from the COLFIS corpus (Laudanna et al., 1995). Imageability of word stimuli was evaluated by ten graduate students in a preliminary study. These student participants were asked to rate each word on a seven-point rating scale that ranged from "very difficult to imagine" to "very easy to imagine," which described the extent to which the concept underlying the word was associated with a mental image. Distribution-based matching was performed because psycholinguistic effects are not always linear (e.g., Bien et al., 2005). The distribution of the variables in the four experimental lists did not differ (Kolmogorov-Smirnov tests: n.s.).

Each experimental list was randomly associated with a list of fillers. The purpose of the filler lists was to prime a prevalent use of either the lexical-semantic reading procedure or the GPC reading procedure. The lexical procedure was elicited through the use of both trisyllabic Italian words and foreign loanwords. The first list of lexical fillers contained 30 trisyllabic words, in which lexical stress was either on the penultimate syllable (15 words, e.g., *parola*, /paˈrola/, word) or on the antepenultimate syllable (15 words, e.g., *tavolo,* /ˈtavolo/, table). The second list of lexical fillers comprised 20 English loanwords (e.g., *barbecue*, /ˈbɑ:bɪkju:/) and 10 French loanwords (e.g., *beige*, /bɛ:ʒ/) that are currently used in Italian but are not readable when following regular GPC rules. GPC reading was elicited by means of pseudowords. Two lists of pseudowords were created to be as orthographically similar as possible to the corresponding lexical filler lists. The first list contained 30 trisyllabic pseudowords with a CV structure (e.g., *dogore*), which matched the filler list of trisyllabic words, and the second list included 30 pseudowords that contained consonant clusters (e.g., *cimpelte*), which matched the filler list of loanwords. The length of pseudowords was matched with the length of familiar words in the corresponding filler lists.

See Appendix 1 for a complete list of the target and filler items.

Experimental Procedure

Target disyllabic words were presented with filler stimuli and presumably elicited a reading process that followed either the lexical-semantic procedure (lexical frame) or the GPC procedure (pseudoword frame). Targets were presented in ten-item blocks that had either a lexical-semantic or a pseudoword frame. The target-word rate was 4/10 for each block. Targets and fillers were presented in semi-randomized order, i.e., in "mini-blocks" that reflected an alternating sequence of 2–3 fillers and 1–2 targets.

Lexical-semantic and pseudoword frame conditions were alternated and counterbalanced across participants. Each frame condition was preceded by a baseline sequence that comprised strings of lines that were oriented differently and were matched with the orthographic stimuli for length, numbers of lines, and visual angle.

There were two separate sessions. In the first session (*loanword-frame session*), the disyllabic target words were embedded in filler lists that were made up of either loanwords or pseudowords that contained consonant clusters (CC). In the second session (*trisyllabic-frame session*), the disyllabic target words were embedded in filler lists that were formed of either trisyllabic Italian words or pseudowords with a CV structure. Participants were randomly assigned to one of the two tasks (**Figure 1**).

All participants in the behavioral study read an additional list of 40 CV-disyllabic target words (block condition).

Therefore, each participant performed only one reading session (a "loanword session" or a "trisyllabic session") because the remaining stimuli were used in the "block condition."

All the stimuli (font: Arial; size: 42; color: black) were displayed in the center of a computer screen on a white background by means of E-Prime software (Psychology Software Tools Inc., Pittsburgh, PA). Participants were instructed to read the letter strings aloud and to press a key on a serial response box for each string of lines. Reading accuracy and voice-onset time (VOT) were recorded. Stimuli remained on the screen until the participant responded. The inter-stimulus interval (ISI) was 1500 ms.

fMRI Study

Participants

A new sample of 20 normal young right-handed adult participants (10 M/12 F; mean age = 24.1 ± 4.4 years) with

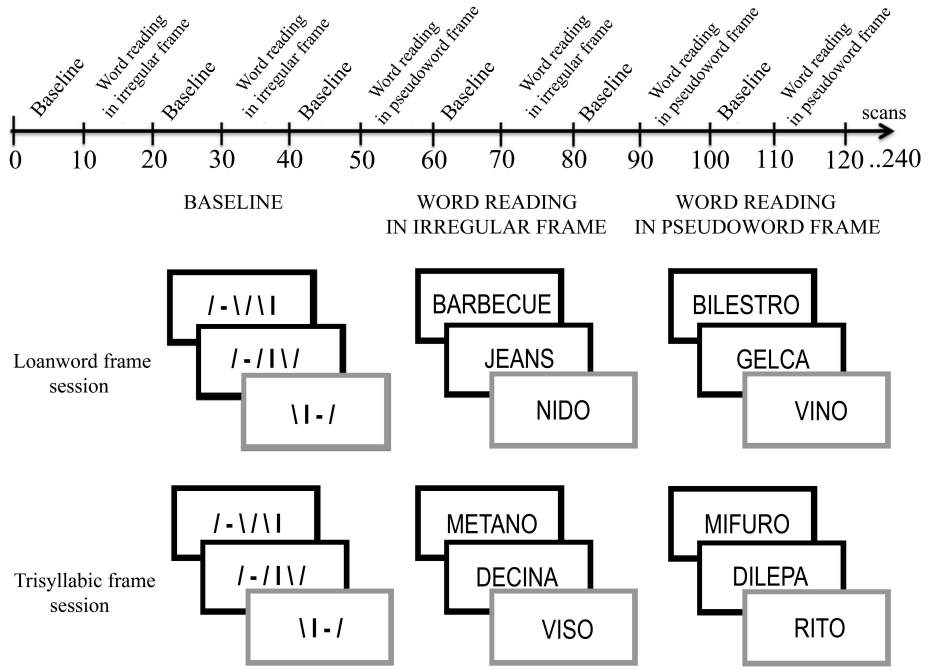


FIGURE 1 | Schematic representation of the time line of tasks.

advanced education (mean education = 15.7 ± 1.7 years) took part in the fMRI study.

All were native Italian speakers with no history of neurological and psychiatric disorder. They all had normal cognitive development, normal or corrected-to-normal vision, and normal language and reading skills.

Informed consent was obtained from all participants prior to the scanning session.

Experimental Procedure

The fMRI study replicated the behavioral experiment as closely as possible.

The fMRI design was based on alternating 30-s baseline blocks (ten blocks for each session) and experimental blocks. In both sessions, the baseline stimuli were strings of lines with different orientation that were matched with the experimental stimuli for length, number of components, and visual angle. The experimental stimuli were the same words, loanwords and pseudowords that were used in the behavioral study (see **Figure 1**).

In the first session *(loanword-frame session)*, the target disyllabic words were alternated with both loanwords (a *loanword frame)* and trisyllabic pseudowords that contained CC (a CC-*pseudoword frame*). In the second fMRI session *(trisyllabic-frame session)*, the disyllabic words were alternated with both trisyllabic Italian words (a *CV-trisyllabic-word frame*) and trisyllabic pseudowords with a CV structure (a *CV-pseudoword frame*). The target-word rate was 4/10 for each block in both sessions.

During the fMRI sessions, stimuli were projected from a PC that was located outside the MR room and was connected by optical fibers to dedicated goggles (Visuastim XGA, Resonance Technology, www.mrivideo.com) using Presentation 11 software (Neurobehavioral Systems, Inc., Albany, CA). Specific corrective lenses were used in the scanner for volunteers with known refraction deficits.

Participants were exposed to each stimulus for 1500 ms in each session. Stimuli were shown in the center of a white screen. The interstimulus interval (ISI) was randomly selected in a time-window of 1200–1800 ms to avoid habituation to the BOLD signal. Participants were exposed to a white screen during the ISI.

Participants were instructed to read words, loanwords, and pseudowords silently to avoid artifacts that would be caused by mouth and head movements. They were asked just to look at the pattern of lines for the baseline task. Participants were also instructed to press a button after each stimulus. Half of the participants pressed the key-button with the right index finger, the other half with the left index finger.

Sessions were presented in a counterbalanced order across participants.

Unlike the behavioral study, all of the participants in the fMRI study performed both the loanword and the trisyllabic sessions, and none performed the block condition.

Image Acquisition

For each participant, 214 fMRI cerebral scans for each reading task were collected using an echo-planar gradient-echo pulse

sequence (EPI; Ogawa et al., 1992), T2* weighted, with a 1.5 T GE-Signa scanner (Slice thickness = 4 mm; Flip angle 90°; TE = 60 ms, TR = 3000 ms, FOV = 240 × 240 mm; matrix = 64 × 64).

fMRI Analyses

The fMRI analyses were performed using the SPM8 software (Welcome Department of Imaging Neuroscience, University College, London). The fMRI images that were collected for all participants were realigned to remove movement artifacts and then were normalized in the MNI-space. Images were then convolved in space with a three-dimensional isotropic Gaussian kernel (10 mm FWHM) to improve the signal-to-noise ratio. A subject-by-task first level analysis was performed after this standard pre-processing step. There were thus two fixed-effect analyses for each participant. The BOLD signal was convolved using the standard hemodynamic response function (HRF) and modeled according to an event-related design (Worsley and Friston, 1995). The event-related matrix was designed to isolate the hemodynamic response that was elicited by reading fillers and by reading targets within each list. Finally, for each subject we estimated the four condition-specific effects of interest: (i) *target reading* > *baseline* in the *loanword frame*, (ii) *target reading* > *baseline* in the *CC-pseudoword frame*, (iii) *target reading* > *baseline* in the *CV-trisyllabic-word frame*, (iv) *target reading* > *baseline* in the *CV-pseudoword frame*. We therefore obtained four "contrast images," i.e., four maps that included the effect of interest per voxel of the brain for each participant. These contrast images were entered into a random-effect analysis that conformed to a general linear model (GLM, Holmes and Friston, 1998; Penny and Holmes, 2004).

A 2*2 second-level ANOVA that had four within-subject conditions (that corresponded to the effects that were described above) was designed and estimated. The factors were *frame session* (loan vs. tri-syllabic) and *lexicality* (word-frame vs. pseudoword-frame).

Because we were interested in disentangling the areas of the reading neural network that are specific to the lexical-semantic procedure from those areas that are specific to the GPC procedure, the ANOVA was explicitly masked by the neural network that was associated with word and pseudoword reading ($p < 0.001$), as described in one of our previous papers (Danelli et al., 2013), i.e., the ANOVA was computed only in the voxels that belonged to the mask.

This allowed us to focus on brain regions that are typically responsive to reading, i.e., the responses should be positive relative to a minimal baseline for word and pseudoword reading, excluding any negative BOLD response within this constrained mask. Moreover, this allowed us to reduce the problem of multiple comparisons (41.0 resels). The mask included the prefrontal and frontal cortex, bilaterally, the insulae, a large part of both temporal lobes, the left parietal lobule, and the primary and secondary visual cortex, bilaterally (see Appendix 2).

We first looked for brain regions that showed a significant interaction effect between experimental condition and frame ($p < 0.001$).

This experimental paradigm helped us to test the neurofunctional correlates of the two reading procedures, while controlling for all psycholinguistic confounds. Indeed, we were comparing the BOLD signal associated with reading "MULO" (mule) with the BOLD signal that is associated with reading "RANA" (frog), i.e., items that are psycholinguistically almost identical.

For this reason, we used both a direct-comparison approach with low threshold ($p < 0.05$) and a less conservative approach based on spatial inference rather than on specific voxel-wise inference. To fully achieve this aim we cleaned the spurious pattern by excluding the activation of "non-interest" by means of an exclusive masking procedure (which has been successfully employed in a number of previous fMRI studies: Pochon et al., 2002; Uncapher et al., 2006; Fliessbach et al., 2007; Danelli et al., 2013).

From the univariate second-level analysis, we extracted:

(1) The direct comparisons between target items in lexical-semantic and sublexical frames (lexical > GPC and GPC > lexical) were computed ($p < 0.05$; spatial threshold = 10 voxels).
(2) Lexical-semantic effect: this was computed as a main effect of the target-word reading in the loanword frame and in the CV-trisyllabic-word frame ($p < 0.001$ uncorrected). This analysis was exclusively masked so that voxels "belonging" to the GPC procedure were excluded from the test. The map for the exclusive mask was generated by using a low threshold ($p < 0.05$ uncorrected). This ensured that the analysis did not consider voxels showing weakest trends for activations in the GPC condition.
(3) GPC effect: this was computed as a main effect of the target reading in the CC-pseudoword frame and in the CV-pseudoword frame ($p < 0.001$ uncorrected). An exclusive masking procedure was used as above but this exclusive mask was derived from the lexical-semantic condition.
(4) Conjunction of the lexical-semantic and GPC effects: the GPC effect and the lexical-semantic effect were entered in a conjunction analysis (Friston et al., 1999; Worsley and Friston, 2000) to identify the brain regions that are commonly activated by both the lexical and the GPC reading procedures ($p < 0.001$). This conjunction was computed as in the univariate analysis adopting a conservative conjunction approach based on minimum statistics procedure (Nichols et al., 2005).

Finally, in order to test whether the isolated networks corresponding to the latest three effects would represent good classifier models of the fMRI images associated with the target reading performance in the lexical or in the sublexical frame, we implemented three multivariate classification analyses, by means of multi-voxel pattern analysis (MVPA), using the PyMVPA 2.2 toolbox (www.pymvpa.org; Hanke et al., 2009). These analyses were implemented in order to support the hypothesis that the exclusive masking could be a valid approach, even if less conservative than a direct comparison method. To this end, we repeated the SPM8 univariate first-level analysis on realigned and spatially normalized, but spatially unsmoothed fMRI data. We computed spmT maps associated with the four condition-specific effects of interest (see the univariate

analysis), which were then used for the MVPA (Misaki et al., 2010).

MVPA was performed on the data of 20 subjects. We trained the linear support vector machine classifier algorithm implemented in PyMVPA with a leave-one-subject-out cross validation procedure, using for each iteration the spmT maps of 19 subjects, and then testing the classification accuracy on the spmT maps (2 for the lexical and 2 for the sublexical condition-specific effects) of the 20th subject. In particular, we ran three different independent multivariate classification analyses using as inclusive mask, respectively, the lexical-semantic, the GPC, and the conjunction effects described above, although at a less conservative significance threshold. These three analyses were run to specifically test the following scenarios:

(1) if the lexical-semantic mask, extracted by means of the massively univariate analysis, actually represented the pool of brain regions exclusively associated with the lexical-semantic reading procedure, then the classifier should be able to accurately distinguish between the lexical-semantic and the GPC spmT maps;
(2) similarly, if the GPC mask extracted from the standard random effect analysis represented the pool of brain regions associated with reading by the GPC procedure, then once again the MVPA should accurately distinguish between the two types of spmT maps;
(3) on the contrary, if the mask extracted from the conjunction effect analysis actually represented the pool of brain regions that are commonly activated by the two procedures, then the MVPA procedure should fail to distinguish between the lexical semantic and the GPC spmT maps.

As for the latter scenario, we further considered whether the MVPA could be a more sensitive approach than the univariate approach, and detect any spatially restricted patterns within the conjunction effect mask, that could distinguish between the lexical semantic and the GPC spmT maps, in spite of a failure at the whole-mask level. To this purpose, we employed recursive feature elimination (Hanson and Halchenko, 2008). Recursive feature elimination was performed strictly within the training partitions, by iteratively eliminating the less sensitive 50% of voxels, and then selecting the reduced brain voxel partition having the greatest sensitivity.

Results

Behavioral Data

The accuracy of all participants was at ceiling for the reading tasks that were performed outside the scanner. VOTs (log-transformed) were analyzed for target words only. Data were trimmed on the basis of the visual inspection of QQ-plots. Datapoints that clearly deviated from a Gaussian distribution (i.e., VOTs that were shorter than 200 ms and longer than 950 ms) were removed.

To account for the non-independence of observations in the dataset, results were analyzed using a mixed-effects model (Baayen et al., 2008) that included random intercepts for items and participants. Outlier datapoints were identified and removed using 2.5 SD of the model residuals as a criterion. Degrees of freedoms were estimated following the method proposed by Satterthwaite (1946).

Data analysis showed a significant main effect of the list (GPC vs. lexical) [$F_{(1,\ 75.31)} = 4.97$; $p = 0.0287$]. Participants were significantly faster when reading disyllabic target words that were embedded in a lexical filler list (mean = 469 ms, SEM = 2.41) than when reading disyllabic target words embedded in a GPC filler list (mean = 482 ms, SEM = 2.38). Neither the interaction between list and task [$F_{(1,\ 75.31)} = 0.69$; $p = 0.4081$], nor the main effect of task [$F_{(1,\ 33.67)} = 0.37$; $p = 0.5442$] were significant.

A second mixed-effects model that also included random slopes for participants was estimated in order to provide indirect evidence that the observed pattern of results (included the absence of behavioral differences between similar frames) depend on participants recruitment biases. The predictions were confirmed: participants were significantly faster [$F_{(1,\ 47.97)} = 7.41$; $p = 0.0089$] when reading disyllabic target words that were embedded in a lexical filler list than when reading disyllabic target words embedded in a GPC filler list, and the interaction between list condition and task was not significant [$F_{(1,\ 47.97)} = 1.14$; $p = 0.2908$]. Indeed, the inclusion of the random slopes did not significantly improve model fit [$X^2_{(6)} = 1.95$; $p = 0.924$], indicating that the associated parameters are not justified by the additional amount of explained variance.

Finally, a further analysis contrasted the responses to the different frame conditions with the responses to the same item in a block design. Participants were significantly faster when reading disyllabic words in a block condition (mean = 458 ms, SEM = 1.75) than in either the lexical [$t_{(132.61)} = 2.42$; $p = 0.0166$] or the GPC filler list [$t_{(133.13)} = 3.24$; $p = 0.0014$].

fMRI Data: Univariate Analyses

No interaction effects emerged from the analyses. This result confirmed the absence of neural differences between either the two lexical frames or the two sublexical frames and justified the evaluation of lexical and sublexical frame effects using t-linear contrasts.

Lexical-semantic Effect

Direct comparison approach (lexical>GPC)

An increased activation was observed in the lexical-semantic frame rather than in the GPC frame at the level of the left hemisphere, and in particular, in the inferior frontal gyrus, bilaterally, in the left precentral and postcentral gyri, in the left superior parietal lobule, in the left superior and middle temporal pole, in the left superior and middle temporal gyrus, in the left hippocampus, in the left inferior occipital gyrus, in the calcarine cortex, in the lingual gyrus and in the cerebellum. Right activations were observed in the superior temporal pole, in the middle temporal gyrus, in the inferior occipital gyrus, in the calcarine cortex and in the cerebellum (**Table 1A**).

Exclusive masking approach

A significant activation was found in the left supplementary motor area (SMA), in the left middle frontal gyrus, in the

TABLE 1 | Brain regions that are significantly activated in direct comparisons between lexical-semantic and sublexical frames ($p < 0.05$; spatial threshold = 10 voxels).

Brain regions	*MNI coordinates*							
	Left hemisphere				Right hemisphere			
	x	*y*	*z*	*Z-score*	*x*	*y*	*z*	*Z-score*
(A) LEXICAL EFFECT > GPC EFFECT								
Inf. frontal gyrus, pars orbitalis	−52	38	−6	3.07				
	−52	40	−2	2.92				
Inf. frontal gyrus, pars triangularis					58	26	2	1.94
Inf. frontal gyrus, pars opercularis	−54	12	6	2.19	52	20	14	2.18
Precentral gyrus	−48	−4	42	2.71				
	−46	−4	46	2.64				
Postcentral gyrus	−54	−14	48	2.94				
	−46	−8	48	2.74				
Sup. parietal lobule	−38	−68	56	3.37	40	18	−22	2.59
Sup. temporal pole	−30	10	−26	3.04				
Mid. temporal pole	−42	16	−26	3.28				
Sup. temporal gyrus	−64	−48	14	2.43				
	−58	−42	14	1.98				
Mid. temporal gyrus	−56	−52	2	2.04	68	−36	4	3.02
	−60	−54	2	1.97	66	−32	2	2.62
Hippocampus	−24	−4	−24	2.81				
Inf. occipital gyrus	−34	−84	−12	2.35	40	−78	−12	2.01
	−36	−88	−8	2.09				
Calcarine cortex	−2	−86	8	2.70	4	−88	10	2.72
	−10	−92	−10	2.10	4	−86	14	2.47
Lingual gyrus	−20	−86	−16	2.15				
	−10	−86	−12	1.94				
Cerebellum	0	−46	−8	3.13	16	−80	−22	3.05
Cerebellum	−14	−84	−18	2.30	22	−82	−24	2.91
(B) GPC EFFECT > LEXICAL EFFECT								
Mid. frontal gyrus, pars orbitalis	−28	44	−12	2.54				
Inf. frontal gyrus, pars orbitalis	−32	36	−8	1.83				
	−36	36	−6	1.82				
Hippocampus	−26	−28	−4	1.98				
Inf. parietal lobule	−44	−40	40	2.27				
	−42	−40	44	2.19				
Fusiform gyrus	−38	−44	−20	2.00				
	−38	−48	−18	1.80				

inferior frontal gyrus, bilaterally, in the left precentral and postcentral gyri, in the left superior parietal lobule, in the left intraparietal sulcus, in the superior temporal pole, bilaterally, in the left superior temporal gyrus, in the middle temporal gyrus, bilaterally, in the left hippocampus, in the left fusiform gyrus, in the left middle occipital gyrus, in the inferior occipital gyrus, bilaterally, in the left V1, in the left lingual gyrus and in the cerebellum, bilaterally (**Table 2A** and areas in blue in **Figure 2**).

GPC Effect

Direct comparison approach (GPC>lexical)

An increased activation was observed in the GPC frame rather than in the lexical-semantic frame at the level of the left middle and inferior frontal gyri, of the left hippocampus, of the left inferior parietal lobule, and of the left fusiform gyrus (**Table 1B**).

Exclusive masking approach

A specific GPC effect was observed in a small subset of left-lateralized brain regions: the middle frontal gyrus, the orbital part of the inferior frontal gyrus, the inferior parietal lobule and the fusiform gyrus. (**Table 2B** and areas in yellow in **Figure 2**).

Conjunction of the Lexical-semantic and GPC Effects

The conjunction analysis revealed shared activation of the middle and inferior frontal gyri, bilaterally, of the SMA, bilaterally, of the left precentral gyrus, of the left inferior parietal lobule, of the

TABLE 2 | Brain regions that are significantly activated in association with either the lexical effect, the GPC effect, or the commonalities between target-word reading in the lexical and target-word reading in the sublexical frames.

Brain regions	*MNI coordinates*							
	Left hemisphere				Right hemisphere			
	x	*y*	*z*	*Z-score*	*x*	*y*	*z*	*Z-score*
(A) LEXICAL EFFECT								
SMA	−10	12	48	3.14				
Mid. frontal gyrus	−44	20	42	3.53				
	−40	10	56	3.25				
Inf. frontal gyrus, pars orbitalis	−50	40	−6	4.47	44	40	−14	2.88
	−52	40	−2	4.42				
Inf. frontal gyrus, pars triangularis					56	24	2	3.62
					58	22	8	3.40
Inf. frontal gyrus, pars opercularis	−56	16	12	3.15	60	14	4	3.55
					60	18	8	3.31
Precentral gyrus	−44	−6	40	3.93				
	−46	−4	36	3.72				
Postcentral gyrus	−52	−12	50	4.39				
	−48	−12	46	4.11				
Sup. parietal lobule	−38	−68	56	4.12				
Intraparietal sulcus	−46	−60	54	3.87				
Sup. temporal pole	−40	18	−24	4.49	44	20	−22	3.92
	−46	12	−20	3.92				
Sup. temporal gyrus	−62	−48	20	3.45				
Mid. temporal gyrus	−62	−48	12	4.05	66	−38	4	3.82
	−56	−44	12	3.73	66	−44	6	3.59
Hippocampus	−24	−4	−24	3.03				
	−26	−8	−24	2.99				
Mid. occipital gyrus	−40	−86	−6	3.17				
Inf. occipital gyrus	−34	−84	−12	3.98	32	−94	−6	3.18
	−32	−88	−8	3.70				
Fusiform gyrus	−32	−80	−14	3.93				
Calcarine cortex	−8	−94	−8	3.53				
	−14	−92	−2	3.50				
Lingual gyrus	−22	−88	−14	3.71				
	−12	−92	−8	3.65				
Cerebellum	−16	−84	−22	3.31	8	−74	−16	4.00
	−30	−74	−28	3.13	30	−64	−28	3.82
Pallidum	−20	0	0	3.00				
	−22	−2	−2	2.93				
(B) GPC EFFECT								
Mid. frontal gyrus	−38	30	30	2.98				
Mid. frontal gyrus, pars orbitalis	−26	46	−14	3.18				
Inf. frontal gyrus, pars orbitalis	−30	44	−16	3.42				
	−36	36	−4	3.22				
Inf. parietal lobule	−48	−42	40	3.28				
Fusiform gyrus	−38	−48	−18	3.25				
(C) CONJUNCTION OF LEXICAL AND GPC EFFECTS								
SMA	−2	10	54	5.26	8	14	52	3.37
Mid. frontal gyrus	−38	46	8	3.61	44	44	26	3.63
	−40	48	12	3.43				
Inf. frontal gyrus, pars orbitalis	−42	20	−6	5.86	50	24	−10	3.38

(Continued)

TABLE 2 | Continued

Brain regions	*MNI coordinates*							
	Left hemisphere				Right hemisphere			
	x	*y*	*z*	*Z-score*	*x*	*y*	*z*	*Z-score*
	−40	32	−2	4.07	48	18	−12	3.34
Inf. frontal gyrus, pars triangularis	−46	34	20	4.27				
	−46	32	16	4.25				
Inf. frontal gyrus, pars opercularis	−44	8	30	3.94				
	−48	12	28	3.92				
Precentral gyrus	−50	8	42	4.84				
	−48	6	46	4.83				
Inf. parietal lobule	−50	−50	42	3.08				
	−52	−46	42	3.02				
Sup. temporal pole	−50	14	−8	6.06				
Fusiform gyrus	−40	−64	−18	4.20				
Mid. occipital gyrus	−18	−102	0	5.80				
Inf. occipital gyrus	−28	−96	−8	4.09	24	−100	−2	4.03
	−32	−92	−10	3.17				
Calcarine cortex					18	−102	0	3.24
					20	−102	4	3.08
Cerebellum	−40	−54	−24	4.68	40	−64	−26	3.18
					34	−70	−28	3.09

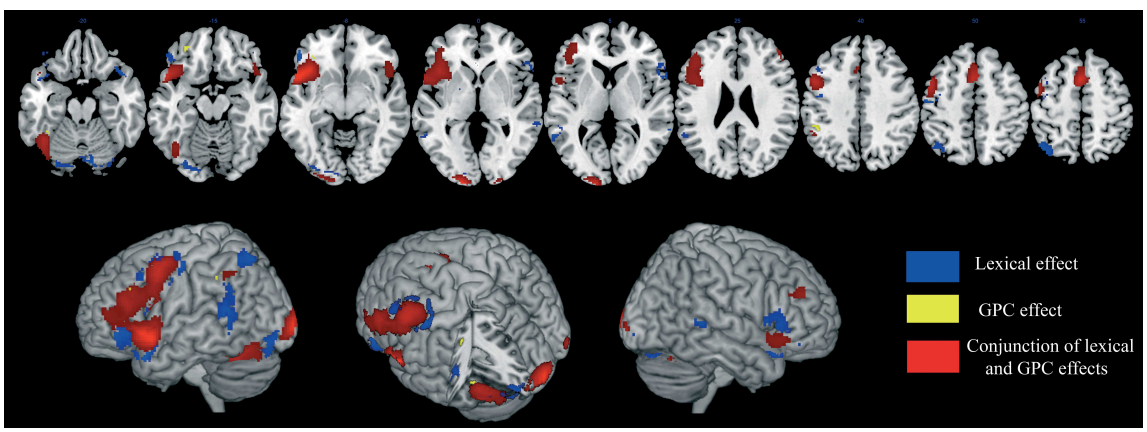


FIGURE 2 | Brain activation data. The cerebral areas that are specifically associated with lexical processing (in blue), sublexical processing (in yellow), and with both reading procedures (in red) are displayed on an anatomical template image (the "ch2better" template image in MRICron; Rorden and Brett, 2000).

left superior temporal pole, of the left fusiform gyrus, and the left middle occipital gyrus, of the inferior occipital gyrus, bilaterally, of the right V1 and of the cerebellum, bilaterally (**Table 2C** and areas in red in **Figure 2**).

fMRI Data: MultiVariate Pattern Analyses (MVPA)

The first multivariate classification analysis (lexical mask) indicated that the neural pattern associated with target reading in the lexical frames represent a valid model (mean classification accuracy = 71.25%; $\chi^2 = 14.5$; $p = 0.002$) to correctly classify the activation patterns associated with target reading in both the lexical (28 out of 40 spmT maps correctly classified) and the sublexical frames (29 out of 40 spmT maps correctly classified).

The second multivariate classification analysis (sublexical mask) showed that the neural pattern associated with target reading in the sublexical frames represented an adequate model (mean classification accuracy = 65.00%; $\chi^2 = 7.4$; $p = 0.06$) to correctly classify the activation patterns associated with target reading in the sublexical frame (27 out of 40 spmT maps correctly classified), but not in the lexical frame (25 out of 40 spmT maps correctly classified).

Finally, the third multivariate classification analysis (conjunction mask) showed that the neural network activated by both the lexical and the sublexical frames did not represent a good model (mean classification accuracy = 51.25%; $\chi^2 = 0.5$; $p = 0.92$) to classify the activation patterns associated with target reading neither in the sublexical frame (19 out 40 spmT maps correctly classified), nor in the lexical frame (22 out 40 spmT maps correctly classified).

Recursive feature elimination further showed that within the conjunction mask there were no spatially restricted activation patterns that could distinguish between the lexical and the sublexical frames (mean classification accuracy = 52.50%; $\chi^2 = 1.0$; $p = 0.80$; 19/40 sublexical and 23/40 lexical spmT maps correctly classified).

Discussion

The neural correlates of single word and pseudoword reading have been investigated in many neuroimaging studies over the past 30 years. These studies identified a left-lateralized cortical network that involved the occipito-temporal cortex, the temporal and temporo-parietal regions and the inferior frontal area, (for reviews, see Fiez and Petersen, 1998; Price, 2012). However, there is only partial agreement on the specific role of these areas in reading, and there is no conclusive evidence that favors a specific model of reading (Bergmann and Wimmer, 2008; Levy et al., 2009; Graves et al., 2010; Roux et al., 2012; Cattinelli et al., 2013).

Our attack on the fortress of the neural correlates of the dual-route model used a list-manipulation paradigm to dissociate the neurofunctional networks that underlie specific (grapheme-to-phoneme conversion, or lexical-semantic) reading procedures and that minimized the effect of stimulus type and task-demand.

We will now discuss the extent to which our behavioral and neurofunctional evidence favors dissociation between the lexical-semantic and GPC reading procedures.

Do the Frames Elicit Prevalent Lexical-semantic rather than Sublexical Reading?

Significant behavioral differences emerged in reading speed between disyllabic words in the irregular word frame and disyllabic words in the pseudoword frame. This result is compatible with lexical-semantic facilitation in one frame, or with a decrement that is related to the prevalent use of the GPC procedure in the pseudoword frame, or with a combination of the two effects.

There would still be disagreement about whether a real facilitation was observed for reading in the lexical frame if one had only the pseudoword frame data as a reference point. However, comparison with an additional baseline measure (disyllabic word reading outside any frame) resulted in the observation that reading lists of target words outside any filler frame is associated with faster reading times. This result can be interpreted in different ways. An explanation might be that participants, when consistently reading the target disyllabic words, become attuned to that word length/orthography while not being *disturbed* by the fillers that come from a different orthography or by trisyllabic words. A more interesting interpretation is that reading the target words in isolation, i.e., outside of the specifically designed filler lists, is accomplished by using all possible strategies, including the sublexical and the lexical-semantic routes. By the same line of reasoning, one can assume that the comparatively longer reaction times for the stimuli in the lexical-semantic filler frame might be due to the prevalent use of a lexical-semantic strategy with relative suppression of the sublexical procedure. This was the effect that we sought with our experimental manipulations.

To summarize, our behavioral results suggest that the word-list manipulations forced participants to emphasize the sublexical GPC procedure in one condition and the lexical-semantic procedure in the other condition. However, this would not lead to reading times that are as fast as those of the "reading-in-isolation" condition, in which participants can let the two non-conflicting procedures (the two horses of the horse-race metaphor, Paap and Noel, 1991) of the dual-route model run freely.

Lexical-semantic and Sublexical Networks: Univariate and Multivariate Analyses

In the fMRI study, we investigated the neural correlates of lexical and sublexical reading procedures using a list-manipulation paradigm. It is worth emphasizing that the BOLD signal was always associated with reading disyllabic words dispersed in the two different frames. To verify whether differences between lexical and sublexical frames may depend on the frame type, a univariate interaction analysis was firstly implemented.

No significant interaction effects emerged from the univariate analysis, suggesting that the functional anatomical differences that are elicited by either the lexical or the sublexical frame should be interpreted as favoring the route-emphasis hypothesis rather than the time-criterion hypothesis. Instead, at a behavioral level, we have no direct evidence for this hypothesis because participants were reading a set of words only in one of the lexical-semantic frame condition, in one of the sublexical condition and in the blocked condition. However, the absence of behavioral differences between similar frames, when random slopes for the participants were included in the mixed-effects model, could provide indirect evidence that the observed effects were not conditioned by a recruitment bias.

Lexical-semantic Procedure

Reading a disyllabic word in a lexical frame activated a specific bilateral set of lexical-semantic regions, specifically the left occipital areas (BA18/19), the posterior part of the middle temporal gyri, the left temporal pole, and the dorsal portion of the left inferior parietal lobule. As confirmed by the multivariate classification analysis, this network represents a valid model to classify the haemodynamic response associated with target reading in both frames. This result supports the hypothesis that these areas are associated with the lexical-semantic procedure of reading.

As reported in literature, the lateral temporal cortex and the posterior portion of the left middle temporal gyrus are often involved in lexical-semantic processing (Vigneau et al., 2006; Binder et al., 2009; Visser et al., 2010). Significant activation of the

left posterior temporal and left parietal regions have indeed been reported for irregular words compared with regular words (Frost et al., 2005; Lee et al., 2005; Senaha et al., 2005), for familiar words compared with pseudowords (Fiebach et al., 2002; Ischebeck et al., 2004; Borowsky et al., 2006), and during semantic tasks compared with phonological decision tasks (McDermott et al., 2003; Mechelli et al., 2005; Booth et al., 2006; see Price, 2012 for a review). Cattinelli et al. (2013) and Taylor et al. (2013) reported the involvement of the left middle temporal cortex in semantic processing that is consistent with these data. In particular, Taylor et al. (2013), in an attempt to clarify the relationship between functional anatomical data of both reading and cognitive models, have suggested that the posterior portion of the left middle temporal gyrus and the angular gyrus would be associated with the phonological lexical and semantic processing.

Our data demonstrate that these cerebral areas are specifically activated during reading through the lexical-semantic procedure and that their activation is independent of such factors as word frequency and imageability. The specific activation of the dorsal portion of the left inferior parietal lobule[3] (together with the activation of the angular and supramarginal gyri), during disyllabic reading in the lexical frame also speaks in favor of an association of this area with the lexical-semantic reading procedure. Partially in line with this result, Taylor et al. (2013) reported a word > pseudoword activation cluster in the left angular and middle temporal gyri, suggesting that this pattern could reflect the engagement (via the orthographic lexicon) of either the phonological lexical or the semantic processing.

Finally, the activation of the left occipital and of the posterior occipito-temporal cortex during disyllabic reading in the lexical frame suggests that there is also preferential processing of words in the early visual areas. This result is consistent with the increased activation observed in the lingual gyrus, which has been interpreted as reflecting the engagement of global shape processing (Mechelli et al., 2000). On the contrary, neither Cattinelli et al. (2013) nor Taylor et al. (2013) observed left occipital- and posterior fusiform-specific activation for words.

Sublexical Procedure

Results of univariate analyses suggest that the left fusiform, the left inferior parietal and the frontal cortex are specifically involved in sublexical reading. As confirmed by the multivariate classification analysis, these cerebral areas, together with the inferior parietal lobule, represent a good model to classify the haemodynamic response associated with target reading in the sublexical frame. This result suggests that these areas are associated with the sublexical reading procedure.

Notwithstanding a little ventral portion of the fusiform gyrus ($x = -38$; $y = -48$; $z = -18$), near to the so-called Visual Word Form Area (VWFA), was activated during disyllabic reading in the sublexical frame, the larger part of this region was activated for reading in both frames (see below for discussion).

[3]The plot of the beta value for each condition, with zero reflecting activity during the baseline condition, showed greater positive activity in the word than in the pseudoword frame within the inferior parietal lobule (−56 −52 38); this suggests the absence of deactivation effects.

Our data also showed that different parietal areas could be associated with different reading procedures. Similar results also emerged in a recent meta-analysis performed by Cattinelli et al. (2013). In particular, our present data show that the left inferior parietal lobule is specifically activated during word reading in the sublexical frame. Consistently with the results obtained by Taylor et al. (2013), the inferior parietal cortex appears to be involved in GPC.

It is worthy to note that the list-manipulation paradigm employed in the present study allowed us to discriminate between the specific neural effects of lexical and sublexical reading and the neural effects that are associated with such linguistic variables as word frequency and imageability, which clearly differ between words and non-words.

Input and Output Components of the Reading Process

The dual-route models that describe the lexical-semantic and sublexical processes as two independent paths predict that some processing units are located upstream and some downstream of the two routes and are shared by both early visual/orthographic input processing and a phonological output buffer. Our results are compatible with this hypothesis. Some brain regions were indeed activated commonly by both the lexical and the sublexical frames. In line with the univariate analyses, the multivariate classification analysis showed that this commonality network does not represent a good model to classify the haemodynamic response associated with target reading either in the lexical or in the sublexical frame. Even spatially more restricted sub-components of the commonality network did not yield successful classification of the lexical vs. the sublexical frames, as shown by recursive feature elimination. Thus, the conjunction brain areas were most likely associated with either early input or late output processes.

With regard to the input visual/orthographic processing in particular, common activation was observed at the level of the left middle occipital cortex and of the left ventral occipito-temporal area, including the so-called Visual Word Form Area (Cohen et al., 2002). Consistent with the dual-route model, the early visual analysis of written words can be described along three steps, which are letter identification, letter position encoding and letter-to-word binding. A deficit in one of these processing stages could cause letter-by-letter dyslexia/pure alexia, which is often associated with a lesion in the left ventral occipito-temporal area (Behrmann et al., 1998; Cuetos and Ellis, 1999; Cohen et al., 2003), and positional dyslexia, which has been associated with a lesion in the occipito-parietal cortex (Friedmann and Gvion, 2001). Additionally, Taylor et al. (2013) reported involvement of the left posterior fusiform and occipito-temporal cortex in non-lexical orthographic processing, which corresponds at a cognitive level to the initial analyses of letter units that are hypothesized by the DRC model. Another interpretation of the activation that emerged in the left ventral occipito-temporal areas for reading in both a lexical and sublexical frame could be termed as an "*orthographic representation matching process*" (Schurz et al., 2010), in which, in the case of words, a visual input is matched with a specific orthographic representation

or in the case of pseudowords, there would be activation of multiple word representations that only partially match visual input[4].

Our data do not allow us to distinguish between these two hypotheses.

Finally, the premotor cortex, the SMA, the left inferior frontal cortex that extend to the anterior part of the left insula and the left prefrontal cortex were commonly activated by both the lexical and the sublexical frames and seem to be associated with the phonological output buffer, i.e., the output store that would support phonological assembling and its interface to covert articulatory plans (see Price, 2012 for a review). In particular, the opercular portion of the left inferior frontal gyrus (LoIFG) is usually considered to be crucial for the reading processes. Neuroimaging studies indicate that the LoIFG is activated more strongly during phonological than during semantic decision tasks for written stimuli (McDermott et al., 2003; Mechelli et al., 2005; Booth et al., 2006), during pseudoword reading than during word reading (Fiebach et al., 2002; Mechelli et al., 2003; Borowsky et al., 2006), and during unfamiliar-word than during familiar-word reading (Fiebach et al., 2002; Ischebeck et al., 2004; Price, 2012). There is convergent evidence from patients with LoIFG lesions, who are impaired in reading pseudowords and low-frequency irregular words (Wagner et al., 2001; Fiez et al., 2006). Cattinelli et al. (2013) suggested that there is an involvement of the LoIFG in more general phonological processing and labeled the LoIFG as an area that is "sensitive to the computational load required by the reading task, rather than to any psycholinguistic variable" and/or processing units (Cattinelli et al., 2013, p. 16). However, while the present data are consistent with the assumption that the LoIFG constitutes a hub of phonological output processes (Taylor et al., 2013), Cattinelli et al.'s (2013) interpretation was further spelled out in terms of difficulty of phonological retrieval in the orthography-to-phonology conversion.

[4]Interestingly, the visual neurocomputational Hmax model (Riesenhuber and Poggio, 2002) for object recognition predicts that a sparse representation (fewer units) should be observed for more finely tuned neural representations, and a non-sparse representation (more units) should be observed for more broadly tuned neural representations. Considering orthographic processing, this model is in line with the assumption that the left occipito-temporal cortex contains neurons tightly tuned to whole words as result of past visual experience with them. Neurons that show high selectivity to a specific word, also show some response to other orthographically similar real words, and to similar pseudowords, thus leading to a total neural signal for pseudowords that might be equal to or even greater than that evoked by real words (Glezer et al., 2009).

Conclusions

The present results provide evidence of shared and divergent neural substrates for the lexical- semantic and the sublexical procedures that underlie word and pseudoword reading. The present results are based on a list-context manipulation and are not confounded by such unbalanced psycholinguistic factors as word frequency and imageability. The list-manipulation procedure may be further exploited to test cross-cultural differences in reading strategies.

It is worthy to note that our study does not provide evidence that favors one particular reading model. Indeed, both the dual-route model and the triangle model predict functional anatomical differences between the two reading frames. Our results showed the existence of a neural dissociation between the lexical and sublexical reading procedures that could be represented by the lexical-semantic and sublexical pathways that are proposed in the dual-route model or by the orthography-to-phonology and the orthography-to-semantics-to-phonology pathways in the triangle model.

Acknowledgments

Preliminary results of this study were presented at the 32nd European Workshop on Cognitive Neuropsychology (Bressanone, Jan. 26–31, 2014), the 20th Congress of Experimental Psychology of the Associazione Italiana di Psicologia (Pavia, Sep. 15–17, 2014), and 52nd Annual meeting of the Academy of Aphasia (Miami, FL, Oct., 5–7, 2014). The study was supported by an FAR grant from the University of Milan-Bicocca and by a grant from Finlombarda-ASTIL 2010 (16873: SAL-20) to CL and EP and from PRIN 2010-2011 (MIUR, prot. 2010ENPRYE_2006) to EP. MM and MB have contributed equally to this work and they are sharing second authorship of the paper; EP and CL have also contributed equally to the study and are sharing senior authorship. All procedures were performed in compliance with relevant laws and institutional guidelines and were approved by the Ethical Committee of the University of Milan-Bicocca (Prot. 124).

References

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

Baluch, B., and Besner, D. (1991). Visual word recognition: evidence for strategic control of lexical and nonlexical routines in oral reading. *J. Exp. Psychol. Learn. Mem. Cogn.* 17, 644–652. doi: 10.1037/0278-7393.17.4.644

Beauvois, M. F., and Derousné, J. (1979). Phonological alexia: three dissociations. *J. Neurol. Neurosurg. Psychiatry* 42, 1115–1124. doi: 10.1136/jnnp.42.12.1115

Bedny, M., Hulbert, J. C., and Thompson-Schill, S. L. (2007). Understanding words in context: the role of Broca's area in word comprehension. *Brain Res.* 1146, 101–114. doi: 10.1016/j.brainres.2006.10.012

Behrmann, M., Nelson, J., and Sekuler, E. B. (1998). Visual complexity in letter-by-letter reading: "pure" alexia is not pure. *Neuropsychologia* 36, 1115–1132. doi: 10.1016/S0028-3932(98)00005-0

Bergmann, J., and Wimmer, H. (2008). A dual-route perspective on poor reading in a regular orthography: evidence from phonological and orthographic lexical decisions. *Cogn. Neuropsychol.* 25, 653–676. doi: 10.1080/02643290802221404

Berlingeri, M., Crepaldi, D., Roberti, R., Scialfa, G., Luzzatti, C., and Paulesu, E. (2008). Nouns and verbs in the brain: grammatical class and task specific effects as revealed by fMRI. *Cogn. Neuropsychol.* 25, 528–558. doi: 10.1080/02643290701674943

Bien, H., Levelt, W. J., and Baayen, R. H. (2005). Frequency effects in compound production. *PNAS* 102, 17876–17881. doi: 10.1073/pnas.0508431102

Binder, J. R., Desai, R. H., Graves, W. W., and Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of

120 functional neuroimaging studies. *Cereb. Cortex* 19, 2767–2796. doi: 10.1093/cercor/bhp055

Booth, J. R., Burman, D. D., Meyer, J. R., Gitelman, D. R., Parrish, T. B., and Mesulam, M. M. (2002). Functional anatomy of intra- and cross-modal lexical tasks. *Neuroimage* 16, 7–22. doi: 10.1006/nimg.2002.1081

Booth, J. R., Lu, D., Burman, D. D., Chou, T. L., Jin, Z., Peng, D. L., et al. (2006). Specialization of phonological and semantic processing in Chinese word reading. *Brain Res.* 1071, 197–207. doi: 10.1016/j.brainres.2005.11.097

Borowsky, R., Cummine, J., Owen, W. J., Friesen, C. K., Shih, F., and Sarty, G. E. (2006). FMRI of ventral and dorsal processing streams in basic reading processes: insular sensitivity to phonology. *Brain Topogr.* 18, 233–239. doi: 10.1007/s10548-006-0001-2

Cappa, S. F., Perani, D., Schnur, T., Tettamanti, M., and Fazio, F. (1998). The effects of semantic category and knowledge type on lexical-semantic access: a PET study. *Neuroimage* 8, 350–359. doi: 10.1006/nimg.1998.0368

Cattinelli, I., Borghese, N. A., Gallucci, M., and Paulesu, E. (2013). Reading the reading brain: a new meta-analysis of functional imaging data on reading. *J. Neurolinguistics* 26, 214–238. doi: 10.1016/j.jneuroling.2012.08.001

Cohen, L., Lehericy, S., Chochon, F., Lemer, C., Rivaud, S., and Dehaene, S. (2002). Language-specific tuning of visual cortex? Functional properties of the visual word form area. *Brain* 125, 1054–1069. doi: 10.1093/brain/awf094

Cohen, L., Martinaud, O., Lemer, C., Lehéricy, S., Samson, Y., Obadia, M., et al. (2003). Visual word recognition in the left and right hemispheres: anatomical and functional correlates of peripheral alexias. *Cereb. Cortex* 13, 1313–1333. doi: 10.1093/cercor/bhg079

Coltheart, M. (1996). Phonological dyslexia: past and future issues. *Cogn. Neuropsychol.* 13, 749–762. doi: 10.1080/026432996381791

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

Coltheart, M., Rastle, K., Perry, C., Langdon, R., and Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychol. Rev.* 108, 204–256. doi: 10.1037/0033-295X.108.1.204

Cuetos, F., and Ellis, A. W. (1999). Visual paralexias in a Spanish-speaking patient with acquired dyslexia: a consequence of visual and semantic impairments? *Cortex* 35, 661–674. doi: 10.1016/S0010-9452(08)70826-8

Danelli, L., Berlingeri, M., Bottini, G., Ferri, F., Vacchi, L., Sberna, M., et al. (2013). Neural intersections of the phonological, visual magnocellular and motor/cerebellar systems in normal readers: implications for imaging studies on dyslexia. *Hum. Brain Mapp.* 34, 2669–2687. doi: 10.1002/hbm.22098

Decker, G., Simpson, G. B., Yates, M., and Locker, L. (2003). Flexible use of lexical and sublexical information in word recognition. *J. Res. Read.* 26, 280–286. doi: 10.1111/1467-9817.00203

Fiebach, C. J., Friederici, A. D., Müller, K., and von Cramon, D. Y. (2002). fMRI evidence for dual routes to the mental lexicon in visual word recognition. *J. Cogn. Neurosci.* 14, 11–23. doi: 10.1162/089892902317205285

Fiez, J. A., Balota, D. A., Raichle, M. E., and Petersen, S. E. (1999). Effects of lexicality, frequency, and spelling-to-sound consistency on the functional anatomy of reading. *Neuron* 24, 205–218. doi: 10.1016/S0896-6273(00)80833-8

Fiez, J. A., and Petersen, S. E. (1998). Neuroimaging studies of word reading. *Proc. Natl. Acad. Sci. U.S.A.* 95, 914–921. doi: 10.1073/pnas.95.3.914

Fiez, J. A., Tranel, D., Seager-Frerichs, D., and Damasio, H. (2006). Specific reading and phonological processing deficits are associated with damage to the left frontal operculum. *Cortex* 42, 624–643. doi: 10.1016/S0010-9452(08)70399-X

Fliessbach, K., Trautner, P., Quesada, C. M., Elger, C. E., and Weber, B. (2007). Cerebellar contributions to episodic memory encoding as revealed by fMRI. *Neuroimage* 35, 1330–1337. doi: 10.1016/j.neuroimage.2007.02.004

Friedman, R. B. (1996). Recovery from deep alexia to phonological alexia: points on a continuum. *Brain Lang.* 52, 114–128. doi: 10.1006/brln.1996.0006

Friedman, R. B., and Kohn, S. E. (1990). Impaired activation of the phonological lexicon: effects upon oral reading. *Brain Lang.* 38, 278–297. doi: 10.1016/0093-934X(90)90115-W

Friedmann, N., and Gvion, A. (2001). Letter position dyslexia. *Cogn. Neuropsychol.* 18, 673–696. doi: 10.1080/02643290143000051

Friston, K. J., Holmes, A. P., Price, C. J., Büchel, C., and Worsley, K. J. (1999). Multisubject fMRI studies and conjunction analyses. *Neuroimage* 10, 385–396. doi: 10.1006/nimg.1999.0484

Frost, S. J., Mencl, W. E., Sandak, R., Moore, D. L., Rueckl, J. G., Katz, L., et al. (2005). A functional magnetic resonance imaging study of the tradeoff between semantics and phonology in reading aloud. *Neuroreport* 16, 621–624. doi: 10.1097/00001756-200504250-00021

Glezer, L. S., Jiang, X., and Riesenhuber, M. (2009). Evidence for highly elective neuronal tuning to whole words in the "Visula Word Form Area." *Neuron* 62, 199–204. doi: 10.1016/j.neuron.2009.03.017

Graves, W. W., Desai, R., Humphries, C., Seidenberg, M. S., and Binder, J. R. (2010). Neural systems for reading aloud: a multiparametric approach. *Cereb. Cortex* 20, 1799–1815. doi: 10.1093/cercor/bhp245

Hagoort, P., Indefrey, P., Brown, C., Herzog, H., Steinmetz, H., and Seitz, R. J. (1999). The neural circuitry involved in the reading of German words and pseudowords: a PET study. *J. Cogn. Neurosci.* 11, 383–398. doi: 10.1162/089892999563490

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., and Pollmann, S. (2009). PyMVPA: a Python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics* 7, 37–53. doi: 10.1007/s12021-008-9041-y

Hanson, S. J., and Halchenko, Y. O. (2008). Brain reading using full brain support vector machines for object recognition: there is no "face" identification area. *Neural Comput.* 20, 486–503. doi: 10.1162/neco.2007.09-06-340

Holmes, A., and Friston, K. J. (1998). Generalisability, random effects and population inference. *Neuroimage* 7, S754.

Ino, T., Nakai, R., Azuma, T., Kimura, T., and Fukuyama, H. (2009). Recognition and reading aloud of kana and kanji word: an fMRI study. *Brain Res. Bull.* 78, 232–239. doi: 10.1016/j.brainresbull.2008.11.008

Ischebeck, A., Indefrey, P., Usui, N., Nose, I., Hellwig, F., and Taira, M. (2004). Reading in a regular orthography: an FMRI study investigating the role of visual familiarity. *J. Cogn. Neurosci.* 16, 727–741. doi: 10.1162/089892904970708

Kang, S. H., Balota, D. A., and Yap, M. J. (2009). Pathway control in visual word processing: converging evidence from recognition memory. *Psychon. Bull. Rev.* 16, 692–698. doi: 10.3758/PBR.16.4.692

Kinoshita, S., and Lupker, S. J. (2007). Switch costs when reading aloud words and nonwords: evidence for shifting route emphasis? *Psychon. Bull. Rev.* 14, 449–454. doi: 10.3758/BF03194087

Laudanna, A., Thornton, A. M., Brown, G., Burani, C., and Marconi, L. (1995). "Un corpus dell'italiano scritto contemporaneo dalla parte del ricevente," in *III Giornate Internazionali di Analisi Statistica dei Dati Testuali,* Vol. 1, eds S. Bolasco, L. Lebart, and A. Salem (Roma: Cisu), 103–109.

Lee, C. Y., Tsai, J. L., Su, E. C.-I., Hung, D. L., and Tzeng, O. J. (2005). Consistency, regularity, and frequency effects in naming Chinese characters. *Lang. Linguist.* 6, 75–107.

Levy, J., Pernet, C., Treserras, S., Boulanouar, K., Aubry, F., Démonet, J. F., et al. (2009). Testing for the dual-route cascade reading model in the brain: an fMRI effective connectivity account of an efficient reading style. *PLoS ONE* 4:e6675. doi: 10.1371/journal.pone.0006675

Lupker, S. J., Brown, P., and Colombo, L. (1997). Strategic control in a naming task: changing routes or changing deadlines? *J. Exp. Psychol. Learn. Mem. Cogn.* 23, 570–590. doi: 10.1037/0278-7393.23.3.570

Marshall, J. C., and Newcombe, F. (1973). Patterns of paralexia: a psycholinguistic approach. *J. Psycholinguist. Res.* 2, 175–199. doi: 10.1007/BF01067101

McDermott, K. B., Petersen, S. E., Watson, J. M., and Ojemann, J. G. (2003). A procedure for identifying regions preferentially activated by attention to semantic and phonological relations using functional magnetic resonance imaging. *Neuropsychologia* 41, 293–303. doi: 10.1016/S0028-3932(02)00162-8

Mechelli, A., Crinion, J. T., Long, S., Friston, K. J., Lambon Ralph, M. A., Patterson, K., et al. (2005). Dissociating reading processes on the basis of neuronal interactions. *J. Cogn. Neurosci.* 17, 1753–1765. doi: 10.1162/089892905774589190

Mechelli, A., Gorno-Tempini, M. L., and Price, C. J. (2003). Neuroimaging studies of word and pseudoword reading: consistencies, inconsistencies, and limitations. *J. Cogn. Neurosci.* 15, 260–271. doi: 10.1162/0898929033212 08196

Mechelli, A., Humphreys, G. W., Mayall, K., Olson, A., and Price, C. J. (2000). Differential effects of word length and visual contrast in the fusiform and lingual gyri during reading. *Proc. Biol. Sci.* 267, 1909–1913. doi: 10.1098/rspb.2000.1229

Misaki, M., Kim, Y., Bandettini, P. A., and Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *Neuroimage* 53, 103–118. doi: 10.1016/j.neuroimage.2010.05.051

Monsell, S., Patterson, K. E., Graham, A., Hughes, C. H., and Milroy, R. (1992). Lexical and sublexical translation of spelling to sound: strategic anticipation of lexical status. *J. Exp. Psychol. Learn. Mem. Cogn.* 18, 452–467. doi: 10.1037/0278-7393.18.3.452

Mummery, C. J., Patterson, K., Hodges, J. R., and Price, C. J. (1998). Functional neuroanatomy of the semantic system: divisible by what? *J. Cogn. Neurosci.* 10, 766–777. doi: 10.1162/089892998563059

Nichols, T., Brett, M., Andersson, J., Wager, T., and Poline, J. B. (2005). Valid conjunction inference with the minimum statistic. *Neuroimage* 25, 653–660. doi: 10.1016/j.neuroimage.2004.12.005

Ogawa, S., Tank, D. W., Menon, R., Ellermann, J. M., Kim, S. G., Merkle, H., et al. (1992). Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proc. Natl. Acad. Sci. U.S.A.* 89, 5951–5955. doi: 10.1073/pnas.89.13.5951

Paap, K. R., and Noel, R. W. (1991). Dual-route models of print to sound: still a good horse race. *Psychol. Res.* 53, 13–24. doi: 10.1007/BF00867328

Paizi, D., Burani, C., De Luca, M., and Zoccolotti, P. (2010). List context manipulation reveals orthographic deficits in Italian readers with developmental dyslexia. *Child Neuropsychol.* 17, 459–482. doi: 10.1080/09297049.2010.551187

Patterson, K., and Behrmann, M. (1997). Frequency and consistency effects in a pure surface dyslexic patient. *J. Exp. Psychol. Hum. Percept. Perform.* 23, 1217–1231. doi: 10.1037/0096-1523.23.4.1217

Patterson, K., and Ralph, M. A. (1999). Selective disorders of reading? *Curr. Opin. Neurobiol.* 9, 235–239. doi: 10.1016/S0959-4388(99)80033-6

Patterson, K., Suzuki, T., and Wydell, T. N. (1996). Interpreting a case of japanese phonological alexia: the key is in phonology. *Cogn. Neuropsychol.* 13, 803–822. doi: 10.1080/026432996381818

Paulesu, E., McCrory, E., Fazio, F., Menoncello, L., Brunswick, N., Cappa, S. F., et al. (2000). A cultural effect on brain function. *Nat. Neurosci.* 3, 91–96. doi: 10.1038/71163

Penny, W., and Holmes, A. P. (2004). "Random-effects analysis," in *Human Brain Function,* eds R. S. J. Frackowiak, J. T. Ashburner, W. D. Penny, S. Zeki, K. J. Friston, C. D. Frith et al. (San Diego, CA: Elsevier), 843–850.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., and Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychol. Rev.* 103, 56–115. doi: 10.1037/0033-295X.103.1.56

Pochon, J. B., Levy, R., Fossati, P., Lehericy, S., Poline, J. B., Pillon, B., et al. (2002). The neural system that bridges reward and cognition in humans: an fMRI study. *Proc. Natl. Acad. Sci. U.S.A.* 99, 5669–5674. doi: 10.1073/pnas.082111099

Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *Neuroimage* 62, 816–847. doi: 10.1016/j.neuroimage.2012.04.062

Price, C. J., Moore, C. J., Humphreys, G. W., and Wise, R. J. (1997). Segregating semantic from phonological processes during reading. *J. Cogn. Neurosci.* 9, 727–733. doi: 10.1162/jocn.1997.9.6.727

Rapcsak, S. Z., Beeson, P. M., Henry, M. L., Leyden, A., Kim, E., Rising, K., et al. (2009). Phonological dyslexia and dysgraphia: cognitive mechanisms and neural substrates. *Cortex* 45, 575–591. doi: 10.1016/j.cortex.2008.04.006

Reynolds, M., and Besner, D. (2005). Contextual control over lexical and sublexical routines when reading english aloud. *Psychon. Bull. Rev.* 12, 113–118. doi: 10.3758/BF03196355

Riesenhuber, M., and Poggio, T. (2002). Neural mechanisms of object recognition. *Curr. Opin. Neurobiol.* 12, 162–168. doi: 10.1016/S0959-4388(02)00304-5

Ripamonti, E., Aggujaro, S., Molteni, F., Zonca, G., Frustaci, M., and Luzzatti, C. (2014). The anatomical foundations of acquired reading disorders: a neuropsychological verification of the dual-route model of reading. *Brain Lang.* 134, 44–67. doi: 10.1016/j.bandl.2014.04.001

Rorden, C., and Brett, M. (2000). Stereotactic display of brain lesions. *Behav. Neurol.* 12, 191–200. doi: 10.1155/2000/421719

Roux, F. E., Durand, J. B., Jucla, M., Réhault, E., Reddy, M., Démonet, J. F. et al. (2012). Segregation of lexical and sub-lexical reading processes in the left perisylvian cortex. *PLoS ONE* 7:e50665. doi: 10.1371/journal.pone.0050665

Rumsey, J. M., Horwitz, B., Donohue, B. C., Nace, K., Maisog, J. M., and Andreason, P. (1997). Phonological and orthographic components of word recognition. A PET-rCBF study. *Brain* 120, 739–759. doi: 10.1093/brain/120.5.739

Sato, H., Patterson, K., Fushimi, T., Maxim, J., and Bryan, K. (2008). Deep dyslexia for kanji and phonological dyslexia for kana: different manifestations from a common source. *Neurocase* 14, 508–524. doi: 10.1080/13554790802372135

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bull.* 2, 110–114. doi: 10.2307/3002019

Schurz, M., Sturm, D., Richlan, F., Kronbichler, M., Ladurner, G., and Wimmer, H. (2010). A dual-route perspective on brain activation in response to visual words: evidence for a length by lexicality interaction in the visual word form area (VWFA). *Neuroimage* 49, 2649–2661. doi: 10.1016/j.neuroimage.2009.10.082

Seidenberg, M. S. (2005). Connectionist models of word reading. *Curr. Dir. Psychol. Sci.* 14, 238–242. doi: 10.1111/j.0963-7214.2005.00372.x

Seidenberg, M. S., and McClelland, J. L. (1989). A distributed, developmental model of visual word recognition and naming. *Psychol. Rev.* 96, 523–568. doi: 10.1037/0033-295X.96.4.523

Senaha, M. L., Martin, M. G., Amaro, E. Jr., Campi, C., and Caramelli, P. (2005). Patterns of cerebral activation during lexical and phonological reading in Portuguese. *Braz. J. Med. Biol. Res.* 38, 1847–1856. doi: 10.1590/S0100-879X2005001200013

Shallice, T., and Warrington, E. K. (1980). "Single and multiple component central dyslexic syndromes," in *Deep Dyslexia*, eds M. Coltheart, K. E. Patterson, and J. C. Marshall (London: Routledge), 326–353.

Tabossi, P., and Laghi, L. (1992). Semantic priming in the pronunciation of words in two writing systems: Italian and English. *Mem. Cognit.* 20, 303–313. doi: 10.3758/BF03199667

Taylor, J. S. H., Rastle, K., and Davis, M. H. (2013). Can cognitive models explain brain activation during word and pseudoword reading? A meta-analysis of 36 neuroimaging studies. *Psychol. Bull.* 139, 766–791. doi: 10.1037/a0030266

Traficante, D., and Burani, C. (2014). List context effects in languages with opaque and transparent orthographies: a challenge for models of reading. *Front. Psychol.* 5:1023. doi: 10.3389/fpsyg.2014.01023

Uncapher, M. R., Otten, L. J., and Rugg, M. D. (2006). Episodic encoding is more than the sum of its parts: an fMRI investigation of multifeatural contextual encoding. *Neuron* 52, 547–556. doi: 10.1016/j.neuron.2006.08.011

Vigneau, M., Beaucousin, V., Hervé, P. Y., Duffau, H., Crivello, F., Houdé, O., et al. (2006). Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing. *Neuroimage* 30, 1414–1432. doi: 10.1016/j.neuroimage.2005.11.002

Visser, M., Jefferies, E., and Lambon Ralph, M. A. (2010). Semantic processing in the anterior temporal lobes: a meta-analysis of the functional neuroimaging literature. *J. Cogn. Neurosci.* 22, 1083–1094. doi: 10.1162/jocn.2009.21309

Wagner, A. D., Maril, A., Bjork, R. A., and Schacter, D. L. (2001). Prefrontal contributions to executive control: fMRI evidence for functional distinctions within lateral Prefrontal cortex. *Neuroimage* 14, 1337–1347. doi: 10.1006/nimg.2001.0936

Wilson, S. M., Brambati, S. M., Henry, R. G., Handwerker, D. A., Agosta, F., Miller, B. L., et al. (2009). The neural basis of surface dyslexia in semantic dementia. *Brain* 132, 71–86. doi: 10.1093/brain/awn300

Worsley, K. J., and Friston, K. J. (1995). Analysis of fMRI time-series revisited–again. *Neuroimage* 2, 173–181. doi: 10.1006/nimg.1995.1023

Worsley, K. J., and Friston, K. J. (2000). A test for a conjunction. *Stat. Probab. Lett.* 47, 135–140. doi: 10.1016/S0167-7152(99)00149-2

Xu, B., Grafman, J., Gaillard, W. D., Ishii, K., Vega-Bermudez, F., Pietrini, P., et al. (2001). Conjoint and extended neural networks for the computation of speech codes: the neural basis of selective impairment in reading words and pseudowords. *Cereb. Cortex* 11, 267–277. doi: 10.1093/cercor/11.3.267

Zevin, J. D., and Balota, D. A. (2000). Priming and attentional control of lexical and sublexical pathways during naming. *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 121–135. doi: 10.1037/0278-7393.26.1.121

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 1

TABLE A1 | List of the written items used in the study.

Lexical frame		Sublexical frame	
Targets	**Fillers**	**Targets**	**Fillers**
SESSION 1			
LAVA	PRIVACY	FETO	VECHENDA
GOLA	LEADER	NASO	SMARILLE
SALA	SHUTTLE	RIGA	CREMPE
MOLO	STEWARD	LUPO	CIRBAGO
RUPE	DISCOUNT	TORO	FOSESCHI
NIDO	TAILLEUR	FILO	CIMPELTE
NANO	COPYRIGHT	FATA	CRUFFELE
RIVA	ZOOM	PIPA	SMETINGA
VENA	POIS	VINO	TENARGE
FUNE	JEANS	VOTO	CHITE
TOPO	TOAST	SETA	GUETA
RISO	FICTION	RAME	GELCA
BUCO	DEEJAY	CANE	OMALIRTO
PELO	BOUQUET	NOCE	COTRENCA
MINA	AUDIENCE	TUTA	APELIARO
PANE	BARBECUE	TIFO	BLACOTA
NAVE	COIFFEUR	LIDO	FASIENE
LAMA	OUTLET	VELA	AOLE
SEME	BOUTIQUE	RANA	DRIMA
MULO	DOWNLOAD	FOCE	ACENPE
	BRIOCHE		CIORGESE
	CROISSANT		FLEDA
	BLACKOUT		BILESTRO
	MOUSE		PRIELI
	YACHT		CLEFO
	BORDEAUX		BURPANTA
	BEIGE		FALTODO
	CLOWN		BRADOLLI
	COMPUTER		SCINEDA
	ROULETTE		ERLA
SESSION 2			
FAVA	MERITO	NODO	COGUNE
ROGO	SEDANO	RUGA	NUSERO
LOBO	TAVOLO	FOTO	SAGATO
FARO	CODICE	PERA	TESOLE
FOCA	BIBITA	FUSO	MIFURO
MELA	FEGATO	SEDE	MAPITO
VISO	REGINA	MIMO	SEFOLO
TUBO	GELATO	VELO	PATOMA
ROSA	VIGILE	LANA	VANOLE
DIVA	DECINA	MURO	FULURA
SALE	CAROTA	TANA	MIRICO
CAVO	METANO	PEPE	SATOME
CORO	PAGINA	MAGO	GETERE
LINO	PATATA	RENE	DOGORE
CODA	NATALE	TELA	POROLO
NEVE	PIRATA	RITO	VIGOTA
SUGO	REGOLA	NUCA	DILEPA
RAMO	CARICA	MUSO	DESARO
LUME	RECITA	FORO	VECATA
VASO	VISITA	DITO	MOPAVO
	COLORE		MUPICA
	LIMITE		CAFEMA
	RESINA		POCORE
	DEBITO		POLEGE
	REGIME		SANUTE
	SENATO		TICOLO
	MINUTO		SEMATA
	MATITA		LISORO
	CUCINA		FEMERA
	RAPINA		CANERA

(Continued)

Appendix 2

TABLE A2 | Brain regions significantly activated during both word and pseudoword reading in Danelli et al. (2013) and used as an explicit mask in the random-effect ANOVA.

Brain regions	*MNI coordinates*							
	Left hemisphere				**Right hemisphere**			
	x	***y***	***z***	***Z-score***	***x***	***y***	***z***	***Z-score***
Sup. frontal med. gyrus	−4	30	52	3.2				
Mid. frontal gyrus					52	14	48	3.3
Inf. frontal orb. gyrus	−44	30	−2	7.0	44	32	−14	5.8
	−40	28	−6	7.0	52	32	−12	5.3
Inf. frontal tri. gyrus					58	28	14	3.9
					62	22	22	4.7
Inf. frontal op. gyrus	−50	16	6	6.1				
	−50	16	20	7.2				
SMA	−6	16	48	4.7				
	−2	6	60	3.4				
Precentral gyrus	−42	2	34	5.7	46	10	48	3.2
	−44	0	54	5.5	48	10	44	3.1
Inf. parietal lobule	−54	−46	54	5.6				
Sup. temporal pole	−50	16	−18	5.8				
Mid. temporal pole					50	16	−26	4.5
Sup. temporal gyrus	−54	−46	20	4.2				
Mid. temporal gyrus	−64	−38	−2	6.5	62	−36	−4	4.8
	−56	−22	−12	4.8	64	−34	−10	4.7
Mid. occipital gyrus	−26	−98	−6	Inf				
Parahippocampal gyrus	−28	−4	−26	4.5				
	−28	−24	−18	4.4				
Inf. occipital gyrus					24	−100	−2	Inf
					32	−90	−10	5.9
Calcarine fissure					4	−82	8	3.2
					8	−84	10	3.2
Vermis					6	−78	−24	4.3
Cerebellum	−44	−54	−24	Inf	34	−76	−22	4.3
	−40	−70	−20	6.6				

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 12 August 2014
doi: 10.3389/fpsyg.2014.00847

Item parameters dissociate between expectation formats: a regression analysis of time-frequency decomposed EEG data

Irene F. Monsalve[1]*, Alejandro Pérez[1] and Nicola Molinaro[1,2]

[1] *BCBL. Basque Center on Cognition, Brain and Language, Donostia, Spain*
[2] *Ikerbasque, Basque Foundation for Science, Bilbao, Spain*

Edited by:
Simona Amenta, University of Milano-Bicocca, Italy

Reviewed by:
Giorgio Arcara, IRCCS, Fondazione Ospedale San Camillo, Italy
Antoine Tremblay, Dalhousie University, Canada
Joost Rommers, Max Planck Institute for Psycholinguistics, Netherlands

****Correspondence:***
Irene F. Monsalve, BCBL. Basque Center on Cognition, Brain and Language, Paseo Mikeletegi 69, 2nd Floor, 20009 Donostia, Spain
e-mail: i.monsalve@bcbl.eu

During language comprehension, semantic contextual information is used to generate expectations about upcoming items. This has been commonly studied through the N400 event-related potential (ERP), as a measure of facilitated lexical retrieval. However, the associative relationships in multi-word expressions (MWE) may enable the generation of a categorical expectation, leading to lexical retrieval *before* target word onset. Processing of the target word would thus reflect a target-identification mechanism, possibly indexed by a P3 ERP component. However, given their time overlap (200–500 ms post-stimulus onset), differentiating between N400/P3 ERP responses (averaged over multiple linguistically variable trials) is problematic. In the present study, we analyzed EEG data from a previous experiment, which compared ERP responses to highly expected words that were placed either in a MWE or a regular non-fixed compositional context, and to low predictability controls. We focused on oscillatory dynamics and regression analyses, in order to dissociate between the two contexts by modeling the electrophysiological response as a function of item-level parameters. A significant interaction between word position and condition was found in the regression model for power in a theta range (~7–9 Hz), providing evidence for the presence of qualitative differences between conditions. Power levels within this band were lower for MWE than compositional contexts when the target word appeared later on in the sentence, confirming that in the former lexical retrieval would have taken place before word onset. On the other hand, gamma-power (~50–70 Hz) was also modulated by predictability of the item in all conditions, which is interpreted as an index of a similar "matching" sub-step for both types of contexts, binding an expected representation and the external input.

Keywords: neuronal oscillations, gamma, theta, anticipatory processes, reading

1. INTRODUCTION

Using previous contextual information in order to anticipate the near future is a pervasive mechanism of human cognition (Bar, 2007), allowing for a fast response to complex stimuli. Such a top-down modulation of perception is also an essential part of language comprehension, where real-time disambiguation involves anticipations about most likely completions. Behavioral studies show that reading times for predictable words are shorter than for unpredictable ones (Ehrlich and Rayner, 1981), demonstrating how prior linguistic context can facilitate linguistic processing. Such predictions may be based on different types of information and occur at different levels. Prior semantic and syntactic content may be used to anticipate a concept and word class that may map onto several lexical items. On the other hand, within certain fixed expressions, a unique word may be unequivocally anticipated, leading to qualitatively different processing.

Previous studies addressing this issue (e.g., Molinaro et al., 2013) have been able to identify differences between compositional contexts and fixed expressions in the event-related potential response (ERP), however, whether this reflects a qualitative difference between the two, or just a stronger expectation in the case of fixed strings remains an open question. The present study aims to address this issue using item-level variability along a number of lexical and orthographic dimensions. Incorporating such item-level variables into the analysis of the electrophysiological response to each type of context will allow a better characterization of the underlying cognitive processes, thus informing neurophysiological models of sentence comprehension.

Within the ERP methodology, the N400 effect is tightly linked with predictability. This ERP component, initially described by Kutas and Hillyard (1980) as a response to semantically anomalous sentence endings, consists of an increased negativity peaking around 400 ms, with a broad scalp distribution. Its amplitude has since then been shown to correlate positively with the predictability of a target word as estimated by its Cloze Probability, (CP[1]: Kutas and Hillyard, 1984), by its word position in the sentence

[1]Percentage of subjects who complete a sentence fragment with a given word.

(Van Petten and Kutas, 1990), and through word probabilities derived from corpus-based models (Frank et al., 2013).

However, the functional interpretation of the N400 component has been debated (e.g., Molinaro et al., 2010). Firstly, an alternative account attributes its modulation not to predictability itself, but to ease of semantic integration (Brown and Hagoort, 1993). Under the predictability view, lexical retrieval would be facilitated through the contextual pre-activation of the given item, whilst under the integration view, facilitation would occur at a combinatorial processing stage, after recognition of the target word had taken place. Federmeier (2007) argues for the predictability interpretation using evidence from previous studies, such as Federmeier and Kutas (1999), that compared the N400 response to unlikely items that had different degrees of semantic similarity to the expected response but would pose similar integration demands (e.g., "They wanted to make the hotel look more like a tropical resort. So, along the driveway, they planted rows of *palms / pines / tulips*). The N400 response was larger for "*tulips*" than for "*pines,*" suggesting that anticipatory activation of "*palms*" would have led to a stronger concurrent activation of "*pines,*" given their shared semantic features (see also Rommers et al., 2013b for a similar paradigm, where anomalous target words sharing only shape-related features with the expected completion also elicited an attenuated N400 response).

Lau et al. (2008) reviewed available evidence from fMRI and MEG localization experiments that employed the same paradigms used in the N400 literature, finding that the only brain region that consistently shows effects under all the reviewed experimental settings is the posterior middle temporal cortex. This is taken as further evidence for the predictive account, given that this area is thought to be involved in lexical/conceptual retrieval, whereas ease or difficulty of integration with prior context should elicit effects in the anterior temporal, inferior parietal and inferior frontal regions.

Semantic constraints may thus facilitate processing at the lexical/conceptual retrieval stage, encompassing, however, a semantic field rather than a specific lexical item. In addition, some studies have been able to show earlier anticipatory effects, acting at the orthographic recognition stage. Kim and Lai (2012) compared the ERP response to semantically constraining sentences where a target word was replaced by an orthographically similar, or dissimilar, pseudoword (e.g., "She measured the flour so she could bake a *cake/ ceke / tont*..."), finding that relative to the expected item (*cake*), the similar pseudoword (*ceke*) elicited a positive deflection at 130 ms, whereas the dissimilar (*tont*) differed from the control, with a different pattern and at a later stage (enhanced negativity at 170 ms). They interpret such an enhanced detection of small, as compared to large, deviations from the target within an interactive top-down/bottom-up framework: when very early bottom-up analysis of the stimulus confirms the top down expectations generated at the conceptual level, further visual analysis stages are enhanced by a specific orthographic prediction. Such an account, whereby conceptual-level expectations percolate down to more specific, visual ones, has also been described at the neural level (Dikker and Pylkkänen, 2013). In an MEG picture-to-word priming task, before the noun was presented, the pictorial contexts elicited activation in left mid-temporal cortex (linked to lexical access), prefrontal cortex (associated with top-down processing) and visual cortex successively.

Nevertheless, early orthographic effects (related to ERP components earlier than the N400) are not as ubiquitous as semantic ones (reflected in the N400). Indeed, anticipating a specific item would in most cases be a difficult task and could lead, overall, to more processing costs than benefits (Jackendoff, 2002). At a semantic level this issue can be resolved by the idea that the expectation, encompassing a set of semantic features, would lead to facilitation of the expected item, but also of its semantic associates. However, given the arbitrary relation between form and meaning in the language system (words such *ant* and *mosquito* are semantically related but not form related), such a semantic neighborhood would not map onto an orthographic one, and pre-activation of the visual features of one word would be of no benefit when processing conceptually similar items. As Kim and Lai's study suggests, only when an initial visual analysis is highly congruent with the orthographic form of the expected item would perceptual top-down facilitation come into place, thus leading to a faster identification of orthographic anomalies.

However, predictions about linguistic stimuli may not be grounded on semantics alone. Associative relationships between words may determine that a specific lexical item, and no other, will appear: such is the case of multi-word expressions (MWE), where particular combinations of lexical items "crystallize" in our semantic memory (Cacciari and Tabossi, 1988, Tremblay et al., 2011). These expressions are pervasive in language, ranging from non-compositional idioms such as "*kick the bucket,*" where the meaning cannot be inferred from the sum of its parts, to collocations, where despite their compositionality, the specific units co-occur with a markedly high frequency, and in a fixed order (such as "*as good as gold,*" or binomials like "*knife and fork*" but not "*fork and knife*:" Siyanova-Chanturia et al., 2011, Arcara et al., 2012).

The ERP correlates to the comprehension of such expressions have been studied by several authors. Roehm et al. (2007a) employed antonym pairs as stimuli, where the second element in the pair was substituted by a same-category or unrelated violation (e.g., "*The opposite of black is white/yellow/nice*"), whilst Vespignani et al. (2010) and Molinaro and Carreiras (2010) used similar paradigms, where MWEs in Italian and Spanish respectively were embedded in sentences where the last item was replaced by a close synonym or a violation[2]. The results of both studies revealed significant graded effects on the N400 amplitude (violation > related item > expected item), but the ERP waveform for the expected completion displayed a particular morphology, with a positive deflection within the initial N400 time-range and a more posterior topography. The authors interpret this as an overlapping P3 response, reflecting the co-occurrence of two qualitatively different processes: a semantic-level anticipation (indexed by the N400), and a partially overlapping categorical target identification mechanism (indexed by the P3). Indeed, the P3b component, a positive deflection peaking around 300 ms with parietal scalp topography, is commonly associated with

[2]The violation condition was only included in the Vespignani et al. study, whilst the Molinaro and Carreiras included additional manipulations.

context updating. In the framework proposed by Kok (2001) it reflects a template-matching process, where an encountered stimulus is compared with an internal representation in a categorical identification process (is it a target or not).

One question that follows from the above studies is whether the hypothesized P3 component arises from the presence of associative relationships between words *per se*, or from the confirmation of a strong expectation that could also be generated by regular compositional contexts. The experimental manipulations consisted of a target-word that was highly expected in one condition (MWE), but unexpected in the others (substitution or violation), so that it is not possible to discern if it was the nature of the expectation or its strength that elicited the results observed. In order to address this question, Molinaro et al. (2013) compared target words that were either embedded in a MWE or in a highly constraining compositional context. By controlling for CP in both conditions, they were able to directly contrast the nature of the predictions: based on associative relationships in one case, and on semantic compositional constraints in the other.

Their results resembled those in previous studies, showing a distinct posterior scalp topography during the first part of the N400 time window (250–350 ms) in the case of MWE, as well as an increased positivity during this same interval that disappeared later on (400–500 ms). The authors interpret these results as support for the presence of two qualitatively different anticipatory processes: a categorical expectation about a specific lexical item (that may either be fulfilled or not), and a graded, semantic expectation (that could be fulfilled to a certain degree). The first process would be more prominent for MWE and the second for highly constraining semantic contexts.

Despite the above experimental results, ERP analysis alone cannot provide conclusive evidence regarding the existence of two qualitatively different cognitive processes during an N400-time window. Firstly, the EEG signal measured at scalp electrodes consists of activity generated by different neuronal populations: if two different sources or networks are active during overlapping intervals, only their summed activity will be recorded at the scalp. Secondly, the ERP averaging process leads to the loss of two kinds of information: (a) any kind of electro-physiological response that despite being time-locked to the stimulus has varying phase across trial (it will be therefore be canceled out through the averaging process); (b) how the effect of interest is modulated by the different lexical and sentence properties of single items.

The Molinaro et al. (2013) study attempted to address some of these limitations by complementing their ERP analysis with oscillatory analysis of EEG phase-locking values (PLV), a method that statistically measures the transient phase coupling between two brain signals in specific frequency bands. Before reading the target word, increased theta phase synchronization was found for the collocational context (over frontal-occipital channels). Furthermore, a positive correlation was found between the increased theta synchronization (before TW onset) and an early post-TW ERP effect (∼120 ms) for the collocational condition only, suggesting that long-range interactions in the theta band support early visual-orthographic analysis of the TW in the case of collocations. However, such PLV results in a pre-TW interval cannot be used to dissociate between the hypothesized P300/N400 overlap.

The present study aims to complement this approach by using regression analysis of the time-frequency decomposition of the data collected by Molinaro et al. (2013) over an N400-like time window. The time-frequency decomposition will provide further information regarding the full dynamics of the EEG response to the stimulus (Makeig et al., 2004), by characterizing the amplitude of oscillations at different frequency bands. The regression analysis will allow the evaluation of whether the frequency characteristics during the time-window of interest (P300/N400 window: 200–600 ms) are influenced by different lexical variables under each condition. Form-based related characteristics, such as the number of orthographic neighbors, may affect the cost of stimulus evaluation and the difficulty of the target-identification task, thus modulating MWE processing (the P3 component is sensitive to both: Herrmann and Knight, 2001). In contrast, lexical and context-related characteristics (such as frequency of use or CP) might be more influential for compositional contexts.

In addition to providing a better characterization of the EEG signal, evidence from the time-frequency domain also has direct functional significance. Increases or decreases in power at certain frequency bands may reflect the dynamic coupling between different brain areas through synchronization of oscillatory activity, thus giving valuable information as to which functional networks become active at different processing stages. Within the language domain, general increases in gamma (>30 Hz) and theta (4–7 Hz) bands, and decreases in alpha (8–12 Hz) ranges have been described in the course of sentence comprehension, with different functional interpretations relating both to predictability and semantic processing (for a review, see Bastiaansen et al., 2012).

Power increases within fast oscillatory activity (gamma-band) can be interpreted as a coupling of near-by neuronal populations arising from successful predictive processing, where representations generated through top-down mechanisms are found to match those generated through bottom-up analysis of the stimulus. Such is the account that Wang et al. (2012) propose for their findings in a study comparing sentences where a target word had either a high CP, low CP, or constituted a semantic violation. They report a parametric modulation of the N400 response (high CP < low CP < semantic violation), but an increase in lower gamma-band power (40–50 Hz; from 0.2 to 1 s post-stimulus onset) over left and posterior electrodes only for the high CP condition. Rommers et al. (2013a) also report increases in gamma power for predictable words in compositional contexts as compared to semantically-related or unrelated violations, albeit over a higher gamma range (50–70 Hz). Interestingly, they also applied the same manipulation to idiomatic contexts, but in this case, no differences in gamma power were found across conditions. Furthermore, a direct comparison between correct compositional and idiomatic expressions revealed higher gamma power for the former in a 60–70 Hz range. They interpret these findings as evidence for the relative "switching off" of semantic operations during idiom comprehension.

Conversely, in a non-sentential paradigm, Dikker and Pylkkänen (2013) found that predictability effects concentrated on the theta band (4–7 Hz), both before and after target word

presentation. They generated predictable or unpredictable contexts for single words using preceding pictures (e.g., picture of an apple vs. picture of a bag with several fruits followed by the word "*apple*"), and examined also the effect of a match or violation for the predictable condition. Before presentation of the target word, more theta band activity for the predictable contexts over left mid-temporal cortex is interpreted as an index lexical retrieval. After target word onset, the contrast between a match or mismatch of the expectation also showed effects in the theta band.

Indeed, results from sentential paradigms also show theta band power increases may accompany lexical retrieval, as well as semantic violations. Bastiaansen et al. (2012) suggest that theta band power increases during lexical retrieval may reflect the binding of semantic properties across distributed representations: the topography of theta-band power accompanying content words was found to be modulated by the semantic properties of the words being processed, so that items with auditory semantic properties elicited theta increases in areas overlying auditory cortex, whilst those with visual semantic properties did so in areas overlying occipital lobes. On the other hand, theta power increases as a result of semantic violations (Davidson and Indefrey, 2007, Wang et al., 2012) could reflect error detection processes.

A complementary view (Klimesch, 1999), attributes theta increases to the encoding of new information, whilst search and retrieval in long-term memory would involve de-synchronization in upper alpha band (∼11–12 Hz), which positively correlated with memory performance. Klimesch related lower alpha band power (∼8–10 Hz) to attentional processes, although the specific boundaries between theta and alpha sub-bands would be subject to high individual variability.

Outside the language domain, Karakaş et al. (2000) studied the decomposition of the P3 ERP component under different paradigms, finding that although it could be explained in terms of superposition of oscillations in lower frequency ranges (delta – 1–3 Hz and theta – 4–7 Hz), a larger amount of variance was explained by delta band oscillations at Pz, with power in this range correlating positively with P3 amplitude. As a result, the delta response is interpreted by the authors as reflecting matching and decision-making operations. Furthermore, Roehm et al. (2007b) re-analyzed the EEG results from the earlier-described antonym study (Roehm et al., 2007a), in order to further dissociate the hypothesized N400/P3 overlapping processes. Indeed, their results showed that the graded N400 response was reflected in qualitative differences in the frequency domain: a delta response (1–3 Hz), maximal at Pz was observed in a comprehension task for the expected antonym pairs only (both in total power and in the time-frequency decomposition of the ERP waveform), but no differences in this range were observed between the two violation conditions. In contrast, a response in the lower theta band (3.5–5 Hz) was reported for the unrelated violation only (although such an increase was not observed in total power, only in the frequency decomposition of the ERP waveform).

Based on the literature reviewed, we could draw the following hypotheses for the present analysis: First, if a categorical, target-identification mechanism is in place during processing of MWE (Roehm et al., 2007a, Molinaro and Carreiras, 2010, Vespignani et al., 2010), a P3-related increase in delta power during the N400 time-window (Karakaş et al., 2000, Roehm et al., 2007b) would be expected for MWEs relative to compositional contexts. A first, low-frequency analysis will therefore focus on the two high CP conditions only. Second, if gamma power increases reflect semantic operations in high predictability contexts (Rommers et al., 2013a), increases in such a power range from 200 ms onwards would be expected when expectations are semantically-based (compositional contexts as compared to low cloze probability controls, Wang et al., 2012), but not when they are based on associative relationships (MWEs as compared to controls), involving a visual, rather than a semantic expectation (Rommers et al., 2013a). However, the specific frequency bands where effects may be detected could be influenced by specific experimental settings and analysis methodologies, so that the whole frequency spectrum will be examined. Finally, if qualitative differences between associative and semantically-based anticipations exist, detected effects could be differently modulated by item-level parameters. Form-based characteristics might be influential for associative anticipations (modulating the difficulty of the target-identification mechanism, Herrmann and Knight, 2001) whilst meaning-based factors could modulate the semantically-based predictions.

2. MATERIALS AND METHODS

2.1. PARTICIPANTS

Thirty-six right-handed native Spanish speakers took part in the experiment (mean age: 22.9, *SD* age: 5.2; 31 females), receiving €10 in exchange for their collaboration. They were all right-handed and had no history of neurological disease. Their vision was normal or corrected to normal.

2.2. MATERIALS

A set of 88 target words (TW) were embedded in three kinds of sentences: collocational contexts, where the TW was the last item in a multi-word expression (MWE)[3]; semantically high-constraining contexts (SEM), where the TW was highly predictable, but not part of a fixed string; and semantically low-constraining sentences (CTR), where the TWs were unpredictable given their previous context, but nevertheless congruent. Target words were the same, and located in the same position within the sentence across conditions at the item level. They were never the last item of the sentence and were always content words. Their cloze-probabilities (as evaluated by an independent group of 40 native Spanish speakers) were very high for the MWE and SEM and did not statistically differ amongst themselves (MWE: Mean: 82.42, *SE*: 2.56; SEM: Mean: 81.56, *SE*: 2.08; $t_{(87)} = 0.27$); CP of TW in the control (CTR) condition was zero.

The MWE used in the first condition were more than three words long (Mean: 4.05, *SE*: 0.10). They were also very frequent expressions, as demonstrated by their frequency of occurrence (Mean: 829.51, *SE*: 215.11) in the *Corpus de Referencia del Español*

[3]These sentences were selected from the stimuli used by Molinaro and Carreiras (2010), which included multi-word expressions extracted from the CESS-ECE corpus (Martí and Taulé, 2007).

Actual (http://corpus.rae.es/creanet.html), and highly familiar, as evaluated through a questionnaire given to 54 independent native Spanish speakers (mean rating: 5.87, *SE*: 0.19, on a 7 point scale where 1: never heard; 7:heard very often). Lexical characteristics (frequency, orthographic neighbors, and length) of the word preceding the target were also controlled for (no t-value larger than 1.32), which was often a function word (CTR: 53; SEM: 48; MWE: 52), and in the remaining cases a content word. This assured that no differences between conditions in the pre-target word interval could derive from the lexical properties of the preceding word, thus minimizing possible uncontrolled carry-over effects. For further details regarding the materials, see Molinaro et al. (2013).

The final experimental set of stimuli was comprised of 264 sentences (see **Table 1** for examples), and an additional 12 sentences used in a practice session.

2.3. PROCEDURE

Participants were tested individually in an electrically-shielded room. Sentences were presented on a CRT computer screen one word at a time. Each word remained on screen for 300 ms and was followed by a 300 ms blank. Yes/No comprehension questions were presented every five sentences on average and sentence order was fully randomized. Twelve practice trials were provided before the experimental session started, which lasted 1 h and 15 min including five breaks across the session. EEG data was simultaneously recorded using BrainAmp system (Brain Products GmbH), through 32 electrodes, at a sampling rate of 500 Hz. Twenty-seven of these were mounted on an EasyCap according to the 10–10 international system (Fp1/2, F3/4, F7/8, Fz, FC1/2, FC5/6, C3/4, Cz, T7/8, CP1/2, CP5/6, P3/4, Pz, P7/8, O1/2), with two electrodes placed on the two mastoid bones and an additional four facial electrodes (two electrodes placed below the two eyes and two electrodes placed on the external chanti of both eyes). Recording was on-line referenced to the left mastoid. Scalp and mastoid electrode impedance was kept below 5 kOhm, and below 10 KOhm for the horizontal eye positions. For further details regarding the procedure, see Molinaro et al. (2013).

Table 1 | Examples of sentence stimuli.

Condition	Example
MWE	Aunque todos éramos incrédulos al respecto, todo se solucionó como por arte de **magia** cuando más falta hacía. *Although we were all skeptical about the issue, everything was solved "as if by art of magic" when it was most needed.*
SEM	El mago nunca revela sus trucos, siempre dice que ha sido cosa de **magia** y no tiene explicación. *The magician never reveals his tricks, he always says it was just magic, and cannot be explained.*
Control	Como estábamos muy estresados Eneko y yo, acudimos anoche a un espectáculo de **magia** y de humor. *Since we were feeling very stressed, Eneko and I went to a magic and humor show last night.*

Target word (TW) appears in bold. English translation for the multi-word expression (quoted values) is literal.

2.4. TIME-FREQUENCY ANALYSIS

Data analysis was carried out in Matlab 2010b, using the FieldTrip toolbox (Oostenveld et al., 2011). EEG was re-referenced offline to the average activity of the two mastoids and filtered with a 0.1–120 Hz band pass filter. The recordings were segmented in time intervals between −1800 and 1000 ms relative to the presentation of the target word. Eye movements, blinks and electrocardiographic artifacts were reduced using independent component analysis (Jung et al., 2000), with subsequent visual inspection of the data to remove any epochs with remaining artifacts. Data from two participants were discarded due to rejection of a high number of trials, and of a further participant due to accidental loss of codes indicating order of trial presentation. From the remaining 33 participants, 6.1% of trials were rejected on average, with no significant across-condition differences [$F_{(2,\ 66)} = 1.11$, $p = 0.3$].

EEG data were then demeaned to eliminate channel bias, by subtracting the mean over the entire epoch from each amplitude value. The time-varying power spectrum of single trials was obtained using two different techniques: a multi-taper approach (Mitra and Pesaran, 1999) for the gamma-range (30–80 Hz) and a Hanning window (500 ms window, 2 Hz frequency steps, 40 ms time steps) for the lower frequencies (0–30 Hz). In the multi-taper analysis, power was calculated using three orthogonal tapers and a time-varying taper length for each frequency (fitting 5 cycles), so that the temporal smoothing decreased with higher frequencies. Time and frequency steps of the sliding window were the same as for the Hanning analysis. Power values were expressed as relative change from a baseline interval calculated from −950 - −650 ms. This is a 300 ms interval prior to the presentation of the word preceding the target (TW-1), rather than the TW itself, which allows direct comparison with the ERP results presented by Molinaro et al. (2013), and minimizes the presence in the baseline of any pre-stimulus predictability effects.

2.5. STATISTICAL ANALYSIS

2.5.1. Confirmatory analysis

Statistical comparisons (for each time, frequency, and electrode over the hypothesized windows) were performed through non-parametric permutation-based t-tests (MWE vs. SEM comparison) and F-tests (involving all three conditions), using 1000 permutations. We hypothesized differences between the two high-expectancy conditions in the delta band, so the two-way comparison (MWE vs. SEM) was used for a low frequency range (1–3 Hz) over an N400-like time window (200–600 ms, Kutas and Federmeier, 2011). On the other hand, we expected differences in the gamma band between low and both types of high CP items, so all conditions were contrasted for a high frequency range (40–70 Hz) encompassing the one described by Wang et al. (2012) and Rommers et al. (2013a).

2.5.2. Exploratory analysis

The above analysis was then extended to include the full frequency range (0–70 Hz), in order to identify other effects not predicted by our hypotheses.

In addition, these comparisons (both for the confirmatory and exploratory analyses) allowed us to further specify the time (ms), frequency (Hz) and space (electrodes) intervals to be considered in the following mixed-effects analysis. Such a selective analysis avoids circularity (Kriegeskorte et al., 2009) by using independent criteria for data selection (differences in the means across conditions) and statistical inference (correlation between power values and several item-level variables). The only predictor in the models that would suffer from circularity is *condition*. Since our selection procedure was based upon differences in condition means, no statistical inferences can be drawn from the presence of a main effect of *condition* in the regression models.

2.5.3. Mixed-effects multiple regression

The log-transformed power averaged over the selected windows served as the dependent variable against which a mixed-effects multiple regression analysis with crossed random effects for subjects and items (Baayen et al., 2008) was performed.

Several item-level variables covering both form-based and meaning-based characteristics of the TWs were included as independent variables in the models (see **Table 2** for descriptive statistics):

- *Number of characters (NRCHAR):* A number of low-level lexical factors, such as number of characters, font type and size, have been reported to affect reading times (Rayner and Pollatsek, 1987). With regard to ERP components, word length affects early stages of processing (∼100 ms), probably reflecting visual analysis of the stimulus, without interacting with the semantic processing of the item (Hauk et al., 2006). Since monospaced fonts were used in the experiment, physical word length could be measured by the number of characters.
- *Orthographic neighbors (NEIGHB):* As with word length, the number of orthographic neighbors (visually similar items, such as *"cat"/"car"*) affects orthographic discrimination of words and can influence both RTs (e.g., McClelland and Rumelhart, 1981) and ERPs (Holcomb et al., 2002). These values, estimated as the Levenshtein distance, were obtained from the *EsPal* database (Duchon et al., 2013).
- *Single word frequency (LOGFREQ):* The effects of word frequency in reading have been repeatedly reported (e.g., Juhasz and Rayner, 2003), although the degree to which this reflects a form-based or meaning-based facilitation derived from familiarity can be questioned (familiarity with the written word-form vs. familiarity with the concept). Baayen (2005) suggests that the tighter correlation of this measure with other word meaning, rather than word form measures, indicates that word frequency mainly indexes conceptual familiarity. Log-transformed word-frequency estimates were obtained from the EsPal database (Duchon et al., 2013).
- *Word bigram frequency (LOGFREQBI):* The frequency of occurrence of two word sequences has also been shown to affect reading times (for a review, see Tremblay, 2012). Log-transformed bigram frequency estimates calculated from bigram counts (CREA corpus) were included in the models in order to control for such effects.
- *Cloze Probability (CLOZEPROB):* The main object of this study is to explore whether the differences in predictive processing of highly predictive compositional vs. associative contexts are qualitative or quantitative. As such, including in the model a measure of predictability allows the estimation of the effects of condition, once quantitative differences in predictability are accounted for. Although cloze probability for our conditions of interest was always high, there was enough variability to allow its inclusion as continuous predictor of power (see **Table 3**). In addition, its values were log transformed to obtain a better spread.
- *Word position (WORDPOS):* Word position in a sentence has been shown to influence RTs, N400 amplitude (Van Petten and Kutas, 1990), and also power estimates over certain frequency bands (Bastiaansen et al., 2002). This has typically been interpreted as a predictability effect: as a sentence develops, higher semantic constraints are placed on upcoming items. This variable was codified as position of the target word from the beginning of the sentence.
- *Trial number (TNUMBER):* Sentence position in the experimental list was included in order to control for fatigue or practice effects.

Initial models included by-subject and by-item intercepts as random effects, and as fixed effects all the item-level variables (centered and scaled) in addition to the interaction of each with a categorical *condition* factor (SEM = 0; MWE = 1). Final models were built by back-fitting fixed effects and forward fitting by-subject and by-item random slopes. First, predictors with $|t| < 2$ were removed one at a time, starting with the interaction terms. The significance of each predictor was assessed through log-likelihood tests, so that only those that improved model fit ($p < 0.05$) were kept in the models. By-subject and by-item random slopes were then assessed individually using likelihood tests.

Table 2 | Item-level variable descriptive statistics.

Variable	Condition	Range	Median	Mean	*SD*
Wordpos	Both	8–24	17.00	17.39	3.23
Nrchar	Both	3–11	5.00	5.45	1.62
Neighbors	Both	1.00–2.60	1.50	1.45	0.39
Logfreq	Both	0.42–3.14	2.15	2.00	0.66
Logfreqbi	MWE	0.09–2.27	0.72	0.89	0.64
	SEM	0.00–2.26	0.78	0.57	0.76
CP	MWE	10–100	92.99	82.22	24.08
	SEM	40–100	90.00	81.56	19.55

Table 3 | Selected windows for mixed-effects analyses.

Window	Time	Frequency	Channels
1. Theta/delta	400–600 ms	2–4 Hz	CP1, CP2, P3, Pz, P4
2. Alpha/theta	260–420 ms	7–9 Hz	F7, F3, FC5, T7
3. Gamma	220–300 ms	50–70 Hz	FC5, T7, CP5, FC1, C3, CP1

Outlier removal was handled after model fitting, since mixed-effect modeling is less vulnerable to extreme values that can critically affect other analyses highly dependent on means aggregation (Baayen and Milin, 2010).

3. RESULTS

3.1. WINDOWS OF INTEREST

Statistical comparisons of the spectral-power estimates were performed using the Resampling Statistical Toolkit, part of the EEGLAB toolbox Delorme and Makeig (2004) for Matlab. The obtained p-values were corrected through the false discovery rate (FDR) method (Benjamini and Yekutieli, 2001), but under this correction, conservative with small effects, no significant differences were found for any of the contrasts in the confirmatory or exploratory analyses. No strong differences between conditions could therefore be detected using averaging-based analysis techniques.

Since the focus of the present study is to use item-level properties to characterize the frequency response in each condition, windows of interest were identified using uncorrected p-values (set at an $\alpha = 0.01$), and subjected to a certain degree of smoothing through inspection of t- and F-maps masked with a more liberal threshold (0.05).

3.1.1. MWE vs. SEM contrast, low frequency bands (0–30 Hz, 0–600 ms post TW)

The t-maps ($p < 0.01$, uncorrected) showed two windows which were selected for further analysis (see **Table 3**):

1. At the boundary between delta and theta bands (2–4 Hz), from 400–600 ms over parietal electrodes (CP1, CP2, P3, Pz, P4). Power over the selected interval was lower for MWE (mean: 1.06, *SE*: 0.03) as compared to SEM (mean: 1.12, *SE*: 0.03).
2. At the boundary between alpha and theta bands (7–9 Hz), from 260–420 ms over left frontal and temporal electrodes (F7, F3, FC5, T7). Power over the selected interval (see **Figures 1**, **2**) was lower for MWE (mean: 0.97 ; *SE*: 0.03) than SEM (mean: 1.08; *SE*: 0.04).

3.1.2. All conditions analysis, high frequency bands (30–70 Hz, 0–600 ms post-TW)

The one-way ANOVA F-maps (contrasting the three conditions MWE, SEM, and CTR) showed differences within an upper gamma band window (50–70 Hz) in the 220–300 ms interval, over left lateralized electrodes (FC5, T7, CP5, FC1, C3, CP1). **Figures 1**, **2** show that power within this frequency during this time-interval is higher for MWE (mean: 1.04, *SE*: 0.01) than SEM (mean: 0.97, $SE = 0.01$), with CTR showing an intermediate pattern (mean: 0.98, $SE = 0.01$).

3.2. MIXED-EFFECTS MODELS

Data was analyzed using the free software statistical package R (R Core Team, 2013) and the *lme4* and *lmerConvenienceFunctions* libraries (Tremblay, 2011; Bates et al., 2012 respectively). Correlations amongst some of the predictors were high, especially between orthographic neighbors and number of characters [$r = 0.77$, $t_{(86)} = 10.18$, $p < 0.001$]. However, multicollinearity diagnostics showed that the problem was not severe (a kappa test on the baseline predictors gave a condition value, κ, of 6.94, indicative of mild co-linearity).

3.2.1. Window 1: Delta/Theta (2–4 Hz)

Neither of the single-item predictors nor their interactions with *condition* were found to be significant.

3.2.2. Window 2: Theta/Alpha (7–9 Hz)

A *condition* by *word-position* interaction was found to be significant by a likelihood test comparing the model with and without the interaction ($\chi^2_{(1)} = 3.83$, $p = 0.05$; see **Table 4** for model coefficients). No by-subject or by-item slopes were significant. Exploration of quartile-quartile plots and residuals revealed normality and homoscedasticity, indicating that the model was

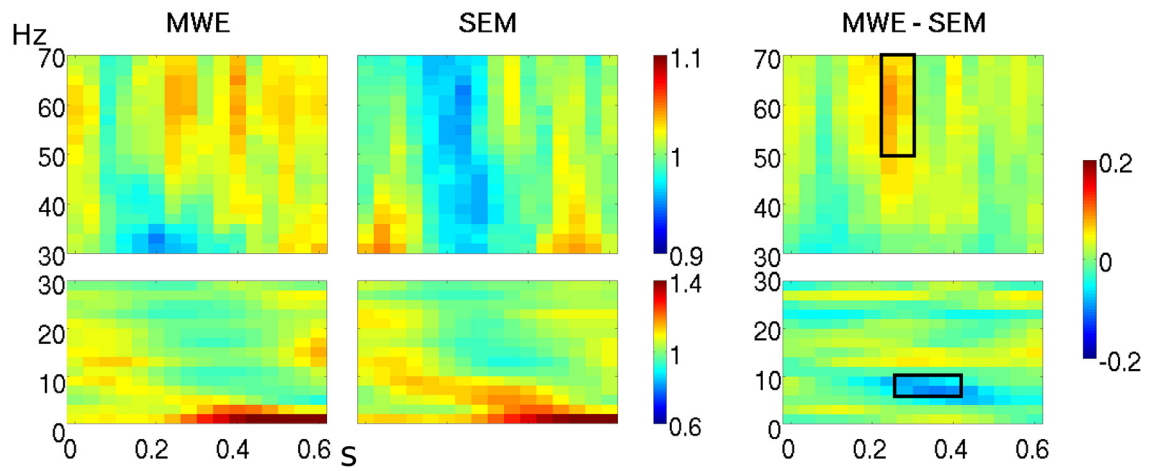


FIGURE 1 | Time-frequency representations of the two high expectancy conditions (MWE, SEM) at electrode T7. High and low frequency ranges are represented separately. The third panel shows the contrast between both conditions, with the selected window for further analysis outlined in black.

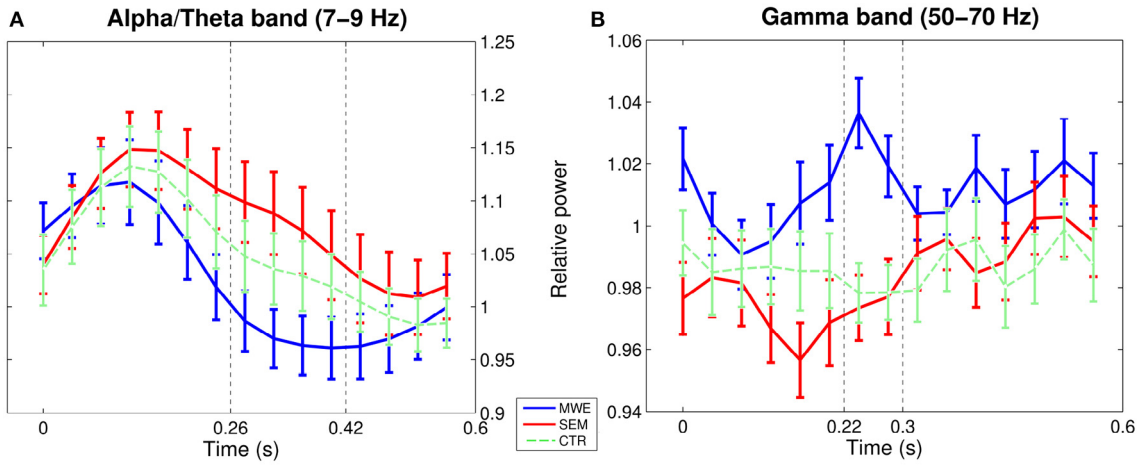


FIGURE 2 | Temporal evolution of power between 0 and 0.6 s post TW onset over selected channel-frequency windows: (A) alpha/theta, (B) mid gamma band. Error bars indicate standard errors around the mean, for illustrative purposes; dotted lines mark analyzed time-window. Although the control condition was not analyzed in the low frequency contrast, it is included in the figure as a reference.

Table 4 | Fixed effects for Theta/Alpha models.

Model	Fixed effects	Estimate	*SE*	*t*-value
MWE coded as 1	(Intercept)	0.230	0.026	8.58
	Wordpos	0.025	0.021	1.16
	Cond	−0.029	0.030	−0.94
	Cond:wordpos	−0.059	0.030	−1.96
SEM coded as 1	(Intercept)	0.197	0.026	7.53
	Wordpos	−0.035	0.021	−1.61

Values for condition and condition-by-word position interaction for both models are the same, and therefore not reported for the second model.

coping well with the data, so no further outlier removal was performed. Variance for the random effects was 0 for by-item intercepts and 0.008 for the by-subject intercepts, with a residual variance of 1.24. Following Barr et al. (2013), we also built a maximal model including by-subject slope for the *condition-by-word-position* interaction (the model did not converge when also including a by-item slope). The fixed effect estimate for the *condition by word position* interaction did not differ from the results reported in **Table 4**, although there was a slight drop in the corresponding *t*-value (−1.92), and in the χ^2 statistic from the likelihood ratio test ($\chi^2_{(1)} = 3.65, p = 0.056$).

The *condition* by *word-position* interaction was tested by fitting an additional model where the *condition* factor was re-coded (MWE = 0, SEM = 1), so that the coefficient for word position reflects the simple slope for each group. The correlation between power and word position was positive for SEM, and negative for MWE, being stronger for the latter (see **Table 4**).

3.2.3. Window 3: Gamma (50–70 Hz)

For the gamma frequency range, *condition* and *cloze probability* remained as predictors in the final model ($|t| > 2$). The significance of *Cloze probability* was confirmed by a log-likelihood ratio test ($\chi^2_{(1)} = 4.33, p = 0.04$). Power levels were higher for the MWE than the SEM condition. Exploration of quartile-quartile plots and residuals revealed that the model was not coping very well with extreme values, deviating from normality. The data was therefore trimmed, eliminating data-points whose residuals were more than 3.5 *SD* away from the mean (29 data points were removed), resulting in a much better model fit (see **Table 5** for trimmed model coefficients). Variance for the random effects was 0.0001 for by-item intercepts and 0.0003 for the by-subject intercepts, with a residual variance of 0.24. Estimates for *cloze probability* fixed effect remained the same after fitting a model with a maximal random effect structure, with χ^2 values obtained through a log-likelihood ratio test dropping slightly ($\chi^2_{(1)} = 3.97$, $p = 0.05$).

4. DISCUSSION

The present study aimed to investigate whether different brain dynamics underlie the predictive response to words embedded either in regular compositional contexts or in MWEs. In the former case, prior semantic information would be used in order to anticipate an upcoming concept and the corresponding likely word candidate. This process, previously linked to the N400 component, would be graded and modulated by the conceptual similarity of the expected item to the actually encountered one. However, several authors (Roehm et al., 2007a, Molinaro and Carreiras, 2010, Vespignani et al., 2010) have proposed that under the associative contexts generated by fixed strings, a categorical expectation is generated, leading to prior lexical retrieval of the upcoming word.

In the case of multi-word expressions, the visual recognition process during reading would thus be akin to a target-identification mechanism, where the encountered stimuli would be compared to an internal representation. Such a process could be indexed by the presence of a P3 effect in comparison to regular compositional contexts. Molinaro et al. (2013) examined this question by comparing MWEs to highly constraining

compositional contexts, finding evidence for the presence of qualitative differences between conditions, through a phase-locking value analysis that revealed differences before presentation of the target word, as well as through an event-related potentials analysis suggesting the presence of a P3 effect for fixed strings. However, the additive nature of the EEG signal and the averaging procedure of the ERP analysis do not allow for conclusive results in this regard. The present study aimed to find further evidence of qualitatively different processes in the post-stimulus interval using time-frequency decomposition of the EEG signal, and regression statistical analyses characterizing the frequency response in terms of item-level variables.

We expected to find differences in two frequency bands: in a delta range, during 200–400 ms and in a gamma range, from 200 ms onwards. Previous research had linked an increase in delta power to target identification mechanisms and the P3 component, during reading of fixed expressions (Roehm et al., 2007b) as compared to compositional contexts, but also in non-linguistic domains (Karakaş et al., 2000). In addition, Wang et al. (2012) reported increases in gamma power during reading of highly expected words as compared to low cloze probability controls, whereas Rommers et al. (2013a) showed that gamma power was higher for semantically constraining contexts than for idiomatic expressions. However, our results revealed no statistically significant differences when comparing power levels averaged over all trials for each condition over the hypothesized time-frequency windows, or over the whole spectrum after correcting for multiple comparisons.

On the one hand, *a priori* determination of frequency bands may miss effects present in the data: small differences between studies employing similar paradigms may lead to substantial differences in the frequency response (see discussions in Klimesch, 1999: regarding individual differences, and Davidson and Indefrey, 2007: regarding the impact of rate of presentation). On the other, statistical comparisons of the full time-frequency-channel data averaged over linguistically variable items may lack the power to detect small effects after correcting for multiple comparisons.

We therefore took an alternative strategy. We used a data-driven approach to select windows of interest (based on maximizing differences between time-frequency-channel data averaged over trials in each condition), and performed a regression analysis to assess how item-level properties modulated the power response (averaged over the selected time-frequency-channels) in each condition, focusing our statistical inference on the latter. This allowed us to evaluate whether both conditions differed in a qualitative way through the presence of condition-by-lexical variable interactions, even when we could not draw inferences regarding differences in the overall means due to the lack of significant results in the selective analysis. Furthermore, the presence of significant main effects of any of the item-level variables may provide information regarding the underlying cognitive processes indexed by power in the given range. In this way, one of the three windows identified (in delta frequency range) was discarded, as no predictors were significant in the mixed-effects model except for condition. We concentrate further discussion on the remaining windows.

Table 5 | Fixed effects for trimmed Gamma model.

Fixed effects	Estimate	*SE*	*t*-value
(Intercept)	−0.061	0.010	−6.07
Cond	0.048	0.014	3.57
Clozeprob	0.014	0.007	2.06

4.1. LOW FREQUENCY RESPONSES

Following the two-way contrast between the semantically constraining sentences and those containing fixed expressions, a cluster at the theta/alpha boundary (6–9 Hz) from 260 to 420 ms over frontal and temporal electrodes in the left hemisphere was selected for further analysis. Overall mean power within this window was lower for MWEs than for compositional contexts (mean: 0.97; *SE*: 0.03 vs mean: 1.08; *SE*: 0.04). However, the regression analysis revealed a condition-by-word-position interaction showing that the differences in power between the two conditions were not constant across the sentence. Theta power was negatively correlated with word position only in the case of fixed strings, and seems to be lower than for compositional contexts only when the target word occurs later on in the sentence, where differences between conditions are maximal (see **Figure 3**).

Such a frequency range, between 6 and 9 Hz could be interpreted as a lower alpha or as a theta effect, given the high inter-individual variability in alpha band frequencies (Klimesch, 1999). Lower-alpha desynchronization has been linked to attentional processes, whilst theta-band synchronization has been linked to lexical-semantic retrieval (Bastiaansen and Hagoort, 2003). However, both the topography (left hemisphere) and the timing of the cluster are more consistent with the language-related theta effects described by Bastiaansen and Hagoort (2003).

Taking theta power to be an index of lexical retrieval, our hypotheses would predict lower power levels for MWEs than compositional contexts: In the case of MWEs only, retrieval of the whole lexical bundle would have taken place at an earlier time-point in the sentence, once the expression is recognized as

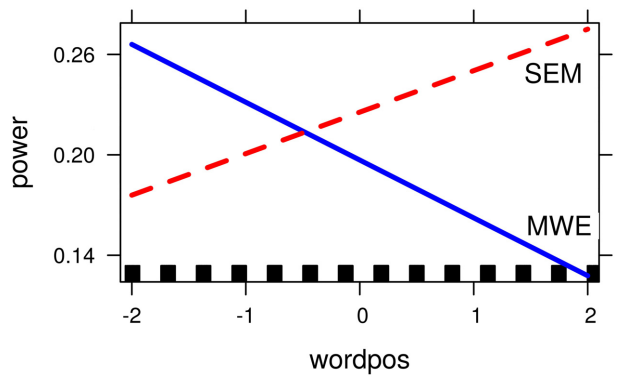


FIGURE 3 | Word-position by condition interaction for Alpha/Theta band model (6–9 Hz). Axis show transformed values for the dependent and independent variable: logarithm for the relative power values, and centered values for word position.

such (recognition point, see Vespignani et al., 2010). In the case of semantically constraining sentences, an anticipatory facilitation could lead to a certain degree of pre-activation, but full retrieval would still require visual recognition of the upcoming item.

However, our results show that the differences in theta power between the two conditions is modulated by target word position, with the expected pattern (lower values for fixed strings) being strongest when the word appears later on in the sentence. If prior lexical retrieval at the recognition point is responsible for differences in theta-band synchronization, it follows that such a recognition point is dependent upon word position in the sentence. The absence of strong semantic constraints at the beginning of a sentence might delay the recognition point to the last element of a fixed expression, so that full retrieval of the lexical bundle would coincide with recognition of the target word. As the sentence unfolds, the increase in contextual semantic information (preceding the onset of the MWE) can lead to an earlier recognition of the fixed expression, allowing for full lexical retrieval of the fixed string before the target word is actually encountered.

We did not find evidence to support our first hypothesis, that predicted P3-related delta increases for fixed strings as compared to compositional contexts. This could be related to differences in the paradigms employed: Roehm et al. (2007b) compared the response to a highly expected antonym with a related-substitution that was nevertheless unexpected. In contrast, in the Molinaro et al. (2013) paradigm both conditions had a high cloze probability. In addition, Roehm et al. showed that the delta response was contingent on the task employed, and could not be detected when it involved lexical decision rather than comprehension. Although the paradigm used in the present study also involved a comprehension task, it differed with the one employed by Roehm et al. in another important aspect: the stimuli included only correct sentences, with no violations.

4.2. HIGH FREQUENCY RESPONSE

Following from the results reported by Wang et al. (2012) and Rommers et al. (2013a), we expected to find predictability-related increases in gamma (40–70 Hz) synchronization from 200 ms onwards (Wang et al.'s effect persists over 1 s) for the semantically constraining contexts as compared to controls and as compared to MWEs. However, our three-way comparison between all conditions revealed no significant differences after correcting for multiple comparisons.

Subsequent window-selection procedure identified a smaller time-window (∼200–300 ms), for a gamma range between 50 and 70 Hz, that was further analyzed using mixed-effects models. Interestingly, the regression model provided evidence that gamma power within this range was indeed related to predictability, with cloze probability being a significant positive predictor of power. There was no significant interaction between this predictor and condition, showing that such a relationship held true across the two high predictability contexts. However, gamma power for the low cloze-probability controls was not lower than for the semantically constraining contexts (mean: 0.98, $SE = 0.01$; mean : 0.97, $SE = 0.01$, respectively). This discrepancy could be explained in terms of differences in the baseline interval used to calculate relative power values. Although the characteristics of words prior to the target were carefully controlled for in Molinaro et al. (2013), cloze probabilities of words preceding the target were considerably lower for controls than for the two high expectancy conditions (see Table 1 in Molinaro et al.). In addition, whether the positive relationship between cloze probability and power held true within the control sentences could not be assessed given the low variability of cloze probability in this condition. For this reason it is critical to evaluate relative differences between the two high expectancy contexts.

Our data is thus consistent with Wang et al.'s (2012) results linking gamma to predictability, but contrary to Rommers et al. (2013a), we cannot link this frequency range to semantically-based anticipations: gamma power was higher for words embedded in idiomatic experessions than for semantically-constraining contexts (see **Figure 2**). Such a discrepancy could be explained in terms of task differences: whilst Rommers et al. used a paradigm that included sentences with expectation violations, our experimental stimuli only contained correct sentences. The proportion of expectation violations in an experimental set has been shown to modulate the the N400 effect (Lau et al., 2013), and cognitive factors like attention Gruber et al. (1999) can modulate gamma-band activity. Attentional patterns may differ in each experimental setting: In a context where only correct sentences are seen an appropriate processing strategy would be to rely on top-down predictions regarding the upcoming word. On the contrary, within the presence of violations more attentional resources may be devoted to bottom-up analysis of the stimulus. If gamma power can be related to predictability across different levels of the cognitive hierarchy, attention-related task differences may modulate at which level (semantic or visual) predictability effects may be enhanced, and therefore detected.

Interestingly, the temporal evolution of power in our case also appears to be different to the one reported by previous studies. Whilst Rommers et al. report gamma synchronization post target word that persists over 1 s for the semantically constraining condition, our results show successive increases and decreases in power values for the two high predictability conditions during the first ∼300 ms, that are nevertheless out of phase, resulting in maximal differences between conditions between 220 and 300 ms (interval that was detected by our data-selection analysis). In contrast, power levels for the control condition remain fairly stable during the whole post-target word interval.

A tentative explanation for such a pattern would be a gamma-rhythm modulation by theta-band oscillations, mechanism that has been proposed to integrate local cell assemblies into large-scale networks (for a review, see Buzsáki and Wang, 2012). Top-down modulation driving the activation of the expected representation would involve large-scale network synchronization in the theta band, whilst successful match with the encountered stimulus could lead to a local increase in gamma-synchronization. Through cross-frequency coupling of gamma power with the theta-rhythm, information about the success of the match may be incorporated into the large-scale network. This process would not be in place for our low predictability sentences, where a successful match is not expected. In addition, the differences in phase between power oscillations for the two high probability conditions could reflect differences in the timing of the predictability

response, with an earlier confirmation of the expectation for the case of MWE. It is important to note, however, that this is only a tentative explanation based on visual inspection of the plots, pointing to an interesting avenue for further analysis of this data-set.

4.3. FINAL REMARKS

In sum, our results provide further evidence of a qualitative difference in anticipatory processing of fixed strings and regular compositional contexts, as evidenced by the differential influence of word position on power in a theta-like range for each type of context. Modeling the frequency response as a function of different item-level variables thus allowed us to better characterize the cognitive processes under each condition, even in the absence of statistically-detectable differences in the overall means.

We suggest that qualitatively different top-down modulation processes in a pre-TW interval could be leading to a pre-activation of certain lexical entries in the case of semantically constraining sentences, and to full retrieval for MWE. Upon encountering the target word, this would lead to subsequent facilitation in lexical retrieval in the former, and a decision to classify the stimulus as a target in the latter. However, the matching step between the bottom-up and the top-down generated representations (whether through full retrieval or pre-activation of an item) would involve the same gamma-band synchronization mechanism, which could show quantitative modulation: earlier in time and with a higher intensity for MWE than compositional contexts. However, our analysis followed an exploratory methodology, so that further research is needed in order to confirm the presented results.

In future studies, we intend to better characterize the different steps of these anticipatory mechanisms, by analyzing a pre-target word interval. It will be interesting to consider how lexical characteristics of the yet-to-come target word influence effects in this time period, and to quantitatively assess cross-frequency coupling. Using MEG and source reconstruction techniques together with individually-determined frequency bands may also enhance the power of the experimental set-up.

Finally, future research into the prevalence and importance of associative relationships between words may bring new insights to our understanding of language function and use. MWEs may play a special role in language, by providing "ready-made" strings to be directly retrieved from memory, thus relieving demands on working memory (Skehan, 1998, Bybee, 2006). The extent to which language relies on such strings, rather than pure compositionality, remains an open question.

ACKNOWLEDGMENTS

This research was partially supported by grants PSI2012-32350. We would like to thank BCBL's Lab Department for the data recording, and especially Larraitz López and Oihana Vadillo, and Margaret Gillon Dowens for reviewing the manuscript.

REFERENCES

Arcara, G., Lacaita, G., Mattaloni, E., Passarini, L., Mondini, S., Benincà, P., et al. (2012). Is hit and run a single word? the processing of irreversible binomials in neglect dyslexia. *Front. Psychol.* 3:11. doi: 10.3389/fpsyg.2012.00011

Baayen, R. H. (2005). "Data mining at the intersection of psychology and linguistics," in *Twenty-First Century Psycholinguistics: Four Cornerstones*, ed A. Cutler (Hillsdale, NJ: Lawrence Erlbaum Associates), 69–83.

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.

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends Cogn. Sci.* 11, 280–289. doi: 10.1016/j.tics.2007.05.005

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

Bastiaansen, M., and Hagoort, P. (2003). Event-induced theta responses as a window on the dynamics of memory. *Cortex* 39, 967–992. doi: 10.1016/S0010-9452(08)70873-6

Bastiaansen, M., Mazaheri, A., and Jensen, O. (2012). "Beyond ERPs: oscillatory neuronal dynamics," in *Oxford Handbook of Event-Related Potential Components*, eds S. Luck and E. Kappenman (New York, NY: Oxford University Press), 31–49.

Bastiaansen, M., van Berkum, J. J., and Hagoort, P. (2002). Event-related theta power increases in the human EEG during online sentence processing. *Neurosci. Lett.* 323, 13–16. doi: 10.1016/S0304-3940(01)02535-6

Bates, D., Maechler, M., and Bolker, B. (2012). *lme4: Linear Mixed-Effects Models Using s4 Classes*.

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29, 1165–1188. Available online at: http://www.jstor.org/stable/2674075

Brown, C., and Hagoort, P. (1993). The processing nature of the n400: evidence from masked priming. *J. Cogn. Neurosci.* 5, 34–44. doi: 10.1162/jocn.1993.5.1.34

Buzsáki, G., and Wang, X.-J. (2012). Mechanisms of gamma oscillations. *Ann. Rev. Neurosci.* 35, 203–225. doi: 10.1146/annurev-neuro-062111-150444

Bybee, J. (2006). From usage to grammar: the mind's response to repetition. *Language* 82, 711–733. doi: 10.1353/lan.2006.0186

Cacciari, C., and Tabossi, P. (1988). The comprehension of idioms. *J. Mem. Lang.* 27, 668–683. doi: 10.1016/0749-596X(88)90014-9

Davidson, D. J., and Indefrey, P. (2007). An inverse relation between event-related and time–frequency violation responses in sentence processing. *Brain Res.* 1158, 81–92. doi: 10.1016/j.brainres.2007.04.082

Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* 134, 9–21. doi: 10.1016/j.jneumeth.2003.10.009

Dikker, S., and Pylkkänen, L. (2013). Predicting language: MEG evidence for lexical preactivation. *Brain Lang.* 127, 55–64. doi: 10.1016/j.bandl.2012.08.004

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., and Carreiras, M. (2013). EsPal: one-stop shopping for Spanish word properties. *Behav. Res. Methods* 45, 1246–1258. doi: 10.3758/s13428-013-0326-1

Ehrlich, S. F., and Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *J. Verbal Learn. Verbal Behav.* 20, 641–655. doi: 10.1016/S0022-5371(81)90220-6

Federmeier, K. D. (2007). Thinking ahead: the role and roots of prediction in language comprehension. *Psychophysiology* 44, 491–505. doi: 10.1111/j.1469-8986.2007.00531.x

Federmeier, K. D., and Kutas, M. (1999). A rose by any other name: long-term memory structure and sentence processing. *J. Mem. Lang.* 41, 469–495. doi: 10.1006/jmla.1999.2660

Frank, S. L., Otten, L. J., Galli, G., and Vigliocco, G. (2013). "Word surprisal predicts N400 amplitude during reading," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Sofia), 878–883.

Gruber, T., Müller, M. M., Keil, A., and Elbert, T. (1999). Selective visual-spatial attention alters induced gamma band responses in the human eeg. *Clin. Neurophysiol.* 110, 2074–2085. doi: 10.1016/S1388-2457(99)00176-5

Hauk, O., Davis, M. H., Ford, M., Pulvermüller, F., and Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *Neuroimage* 30, 1383–1400. doi: 10.1016/j.neuroimage.2005.11.048

Herrmann, C. S., and Knight, R. T. (2001). Mechanisms of human attention: event-related potentials and oscillations. *Neurosci. Biobehav. Rev.* 25, 465–476. doi: 10.1016/S0149-7634(01)00027-6

Holcomb, P. J., Grainger, J., and O'rourke, T. (2002). An electrophysiological study of the effects of orthographic neighborhood size on printed word perception. *J. Cogn. Neurosci.* 14, 938–950. doi: 10.1162/089892902760191153

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. New York, NY: Oxford University Press.

Juhasz, B. J., and Rayner, K. (2003). Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *J. Exp. Psychol. Learn. Mem. Cogn.* 29, 1312. doi: 10.1037/0278-7393.29.6.1312

Jung, T.-P., Makeig, S., Humphries, C., Lee, T.-W., Mckeown, M. J., Iragui, V., et al. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology* 37, 163–178. doi: 10.1111/1469-8986.3720163

Karakaş, S., Erzengin, Ö. U., and Başar, E. (2000). A new strategy involving multiple cognitive paradigms demonstrates that erp components are determined by the superposition of oscillatory responses. *Clin. Neurophysiol.* 111, 1719–1732. doi: 10.1016/S1388-2457(00)00418-1

Kim, A., and Lai, V. (2012). Rapid interactions between lexical semantic and word form analysis during word recognition in context: evidence from ERPs. *J. Cogn. Neurosci.* 24, 1104–1112. doi: 10.1162/jocn-a-00148

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

Kok, A. (2001). On the utility of P3 amplitude as a measure of processing capacity. *Psychophysiology* 38, 557–577. doi: 10.1017/S0048577201990559

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., and Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neurosci.* 12, 535–540. doi: 10.1038/nn.2303

Kutas, M., and Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp). *Ann. Rev. Psychol.* 62, 621–647. doi: 10.1146/annurev.psych.093008.131123

Kutas, M., and Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science* 207, 203–205. doi: 10.1126/science.7350657

Kutas, M., and Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature* 307, 161–163. doi: 10.1038/307161a0

Lau, E. F., Holcomb, P. J., and Kuperberg, G. R. (2013). Dissociating n400 effects of prediction from association in single-word contexts. *J. Cogn. Neurosci.* 25, 484–502. doi: 10.1162/jocn-a-00328

Lau, E. F., Phillips, C., and Poeppel, D. (2008). A cortical network for semantics:(de) constructing the N400. *Nat. Rev. Neurosci.* 9, 920–933. doi: 10.1038/nrn2532

Makeig, S., Debener, S., Onton, J., and Delorme, A. (2004). Mining event-related brain dynamics. *Trends Cogn. Sci.* 8, 204–210. doi: 10.1016/j.tics.2004.03.008

Martí, M. A., and Taulé, M. (2007). *CESS-ECE: Corpus Anotados Del Español y Catalán*. Bergen: Arena Romanística: Corpus and text linguistics in Romance languages.

McClelland, J. L., and Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychol. Rev.* 88, 375.

Mitra, P. P., and Pesaran, B. (1999). Analysis of dynamic brain imaging data. *Biophys. J.* 76, 691–708. doi: 10.1016/S0006-3495(99)77236-X

Molinaro, N., Barraza, P., and Carreiras, M. (2013). Long-range neural synchronization supports fast and efficient reading: EEG correlates of processing expected words in sentences. *Neuroimage* 72, 120–132. doi: 10.1016/j.neuroimage.2013.01.031

Molinaro, N., and Carreiras, M. (2010). Electrophysiological evidence of interaction between contextual expectation and semantic integration during the processing of collocations. *Biol. Psychol.* 83, 176–190. doi: 10.1016/j.biopsycho.2009.12.006

Molinaro, N., Conrad, M., Barber, H. A., and Carreiras, M. (2010). On the functional nature of the N400: contrasting effects related to visual word recognition and contextual semantic integration. *Cogn. Neurosci.* 1, 1–7. doi: 10.1080/17588920903373952

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.-M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* 2011, 1. doi: 10.1155/2011/156869

Rayner, K., and Pollatsek, A. (1987). "Eye movements in reading: a tutorial review," in *Attention and Performance XII: The Psychology of Reading*, Vol. 12, ed M. Coltheart (London: Erlbaum), 327–362.

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org/

Roehm, D., Bornkessel-Schlesewsky, I., Rösler, F., and Schlesewsky, M. (2007a). To predict or not to predict: influences of task and strategy on the processing of semantic relations. *J. Cogn. Neurosci.* 19, 1259–1274. doi: 10.1162/jocn.2007.19.8.1259

Roehm, D., Bornkessel-Schlesewsky, I., and Schlesewsky, M. (2007b). The internal structure of the N400: frequency characteristics of a language-related ERP component. *Chaos Complex. Lett.* 2, 365–395.

Rommers, J., Dijkstra, T., and Bastiaansen, M. (2013a). Context-dependent semantic processing in the human brain: evidence from idiom comprehension. *J. Cogn. Neurosci.* 25, 762–776. doi: 10.1162/jocn-a-00337

Rommers, J., Meyer, A. S., Praamstra, P., and Huettig, F. (2013b). The contents of predictions in sentence comprehension: activation of the shape of objects before they are referred to. *Neuropsychologia* 51, 437–447. doi: 10.1016/j.neuropsychologia.2012.12.002

Siyanova-Chanturia, A., Conklin, K., and van Heuven, W. J. (2011). Seeing a phrase time and again matters: the role of phrasal frequency in the processing of multiword sequences. *J. Exp. Psychol. Learn. Mem. Cogn.* 37, 776. doi: 10.1037/a0022531

Skehan, P. (1998). *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press.

Tremblay, A. (2011). Lmerconveniencefunctions: a suite of functions to back-fit fixed effects and forward-fit random effects.

Tremblay, A. (2012). Empirical evidence for an inflationist lexicon. *Yearbook Phraseol.* 3, 109–126. doi: 10.1515/phras-2012-0006

Tremblay, A., Derwing, B., Libben, G., and Westbury, C. (2011). Processing advantages of lexical bundles: evidence from self-paced reading and sentence recall tasks. *Lang. Learn.* 61, 569–613. doi: 10.1111/j.1467-9922.2010.00622.x

Van Petten, C., and Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Mem. Cogn.* 18, 380–393.

Vespignani, F., Canal, P., Molinaro, N., Fonda, S., and Cacciari, C. (2010). Predictive mechanisms in idiom comprehension. *J. Cogn. Neurosci.* 22, 1682–1700. doi: 10.1162/jocn.2009.21293

Wang, L., Zhu, Z., and Bastiaansen, M. (2012). Integration or predictability? A further specification of the functional role of gamma oscillations in language comprehension. *Front. Psychol.* 3:187. doi: 10.3389/fpsyg.2012.00187

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.

Received: 27 February 2014; accepted: 16 July 2014; published online: 12 August 2014.
Citation: Monsalve IF, Pérez A and Molinaro N (2014) Item parameters dissociate between expectation formats: a regression analysis of time-frequency decomposed EEG data. Front. Psychol. ***5****:847. doi: 10.3389/fpsyg.2014.00847*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in Psychology

ORIGINAL RESEARCH
published: 16 April 2015
doi: 10.3389/fpsyg.2015.00452

How language affects children's use of derivational morphology in visual word and pseudoword processing: evidence from a cross-language study

Séverine Casalis[1]*, *Pauline Quémart*[2] *and Lynne G. Duncan*[3]

[1] SCALab, Université de Lille and Centre National de la Recherche Scientifique, Villeneuve d'Ascq, France, [2] University of Poitiers and Centre National de la Recherche Scientifique, Poitiers, France, [3] School of Psychology, University of Dundee, Dundee, UK

OPEN ACCESS

Edited by:
Simona Amenta,
University of Milano-Bicocca, Italy

Reviewed by:
Cristina Burani,
Institute of Cognitive Sciences and Technologies-Consiglio Nazionale delle Ricerche, Italy
John Kirby,
Queen's University, Canada

***Correspondence:**
Séverine Casalis,
SCALab, Université de Lille and Centre National de la Recherche Scientifique, (UMR 9193), Rue du barreau, 59653 Villeneuve d'Ascq, France
severine.casalis@univ-lille3.fr

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

Received: 01 September 2014
Accepted: 30 March 2015
Published: 16 April 2015

Citation:
Casalis S, Quémart P and Duncan LG (2015) How language affects children's use of derivational morphology in visual word and pseudoword processing: evidence from a cross-language study. Front. Psychol. 6:452. doi: 10.3389/fpsyg.2015.00452

Developing readers have been shown to rely on morphemes in visual word recognition across several naming, lexical decision and priming experiments. However, the impact of morphology in reading is not consistent across studies with differing results emerging not only between but also within writing systems. Here, we report a cross-language experiment involving the English and French languages, which aims to compare directly the impact of morphology in word recognition in the two languages. Monolingual French-speaking and English-speaking children matched for grade level (Part 1) and for age (Part 2) participated in the study. Two lexical decision tasks (one in French, one in English) featured words and pseudowords with exactly the same structure in each language. The presence of a root (R+) and a suffix ending (S+) was manipulated orthogonally, leading to four possible combinations in words (R+S+: e.g., postal; R+S−: e.g., turnip; R−S+: e.g., rascal; and R-S-: e.g., bishop) and in pseudowords (R+S+: e.g., pondal; R+S−: e.g., curlip; R−S+: e.g., vosnal; and R−S−: e.g., hethop). Results indicate that the presence of morphemes facilitates children's recognition of words and impedes their ability to reject pseudowords in both languages. Nevertheless, effects extend across accuracy and latencies in French but are restricted to accuracy in English, suggesting a higher degree of morphological processing efficiency in French. We argue that the inconsistencies found between languages emphasize the need for developmental models of word recognition to integrate a morpheme level whose elaboration is tuned by the productivity and transparency of the derivational system.

Keywords: morphology, reading acquisition, cross language comparison, visual word recognition, lexical decision task

Introduction

In a recent paper, Frost (2012) has put forward a case for a universal model of reading. As it is not certain that, as a cultural product, written language should be subject to a universal form of processing (Coltheart and Crain, 2012), it seems important to consider whether variations in language

properties constrain the use of particular language units. Deacon (2012) rightly pointed out the relevance of the developmental approach to deal with this issue since a key aspect of developmental studies is that they tell us which skills drive reading acquisition and which are the product of reading. In other words, developmental cross language studies should help to disentangle which aspects of reading acquisition are universal and which depend on language properties. Therefore, the aim of the present paper is to compare how English-speaking and French-speaking developing readers make use of one of the fundamental units of reading development, namely, morphemes.

Research conducted over three decades has documented the importance of phonological coding in the earliest phases of learning to read an alphabetic script (Goswami and Bryant, 1990; Muter et al., 2004; Melby-Lervåg et al., 2012). The key unit relevant for phonological coding in alphabetic scripts is the grapheme and its oral counterpart, the phoneme. However, spelling-to-sound consistency varies across orthographies (Frost et al., 1987) and alphabetic orthographies are distinguished along a continuum from transparent to opaque. In some orthographies, grapheme-phoneme correspondences (hereafter, GPC) are transparent, with individual graphemes always pronounced in the same way (e.g., Finnish, Italian). In other orthographies, GPC are highly opaque, meaning that the same grapheme can be pronounced in different ways (e.g., English). The French orthography is opaque in terms of spelling, indeed similar to English in this respect, but more transparent when it comes to reading.

Orthographic transparency has an impact on the ease with which children learn to read across countries, and reading achievement at the end of the first year clearly depends on the consistency of the GPC (Seymour et al., 2003; Duncan et al., 2013). Learning to read is particularly difficult for English-speaking children, who perform at a much lower level than children learning to read in other languages. This delay is observed when reading both familiar words and pseudowords in the initial phases of reading acquisition (Seymour et al., 2003). The Psycholinguistic Grain-Size theory (PGST, Ziegler and Goswami, 2005) has been proposed to account for the effects of such cross-language differences in orthographic depth on reading acquisition (Ziegler et al., 2001; Ziegler and Goswami, 2006). This model suggests that reading development across alphabetic scripts may display some variation in the grain size that children utilize as a function of the availability of units in oral language and the consistency of the links between these units of speech and written orthographic symbols. Even though the PGST focuses on reading aloud, some features may generalize to other aspects of reading such as silent visual word recognition. Equally, other written units such as morphemes, which are not included in this model may come to play a role during literacy acquisition, particularly if these units are available in language and resolve irregularity within the orthography (Ziegler et al., 1997).

The role of morphology in learning to read alphabetic scripts has received increased attention over the past two decades due to a number of factors: (1) most alphabetic writing systems are morphophonemic, in that they represent both phonemic and morphemic units; (2) the majority of new words that children encounter in print are morphologically complex (Nagy and Anderson, 1984), which means that decomposing complex words into smaller constituents during visual word recognition should be particularly relevant when learning to read these words; and (3) developing readers have acquired morphological awareness of spoken language and represent morphological information within their lexicon (Duncan et al., 2009), so given the "intimate relationship between spoken and written language skills" (Hulme and Snowling, 2013, p. 1), word reading is likely to draw upon this ability, particularly in the case of morphologically complex words.

The role of morphology in children's visual word processing has been examined across several languages. In English, children name derived words (e.g., *dancer*) more accurately than pseudoderived words (e.g., *dinner*) as early as Grade 2 (Laxon et al., 1992; Carlisle and Stone, 2005). This effect depends on family size, i.e. the number of derived forms (Carlisle and Katz, 2006), and on base word frequency for reading accuracy (Mann and Singson, 2003; Carlisle and Stone, 2005) and reading speed (Deacon et al., 2011). In Italian, third and fifth graders read pseudowords made up of morphemes (e.g., *donnista*) faster than control pseudowords (e.g., *donnosto*) (Burani et al., 2002, 2008). In relation to words, Italian children read derived words faster than non-derived words but this effect is limited to low frequency words (Marcolini et al., 2011). Finally, the presence of a base and/or a suffix facilitates visual word recognition in the French language (Quémart et al., 2012). When combined, such units also slow down lexical decisions, give rise to a high false alarm rate (Quémart et al., 2012) and enhance speed and accuracy of pseudoword naming (Colé et al., 2012).

Together, these results strongly support the importance of morphemes for developing readers when reading and/or accessing the lexicon but fail to provide a unified picture of the conditions under which this facilitation occurs, since the effects of morphological structure were not consistent. First, morphological structure significantly influenced both accuracy and latencies in French but was significant for accuracy only in English. Second, morphemes affected reading and lexical access when embedded in words and pseudowords in French, whereas such effects were observed only when morphemes were located within words in English and within pseudowords in Italian, except when words were low in frequency. Third, grade level or age of the participants was not constant across studies and there is reason to believe that the contribution of morphology to word processing is not the same during the first steps of reading acquisition as it is later when decoding mechanisms are well developed and more automatic. Finally, at least two different tasks have been used in previous studies, naming and lexical decision, complicating comparisons. Thus, to shed light on how language affects the use of morphology, cross-language studies using equivalent stimuli, a similar procedure and children at comparable grade or age levels are necessary.

To achieve this goal, the present study compares sensitivity to morphemes during visual word recognition among children speaking French vs. English. The French language is acknowledged as a morphologically rich language, with approximately 75% of French words being morphologically complex (Rey-Debove, 1984), while in English, morphologically complex

(derived) forms account for 55% of the lemmas in the CELEX English database. Compounding is more prevalent in English than French, and is not especially productive in French especially in colloquial speech, thus word formation relies far more on derivation than compounding (Clark, 1998; Bauer, 2003). Indeed, French children perform higher in derived form production than English children (Duncan et al., 2009). In the present study, therefore, two effects may act in opposing directions on the outcome: first, the higher prevalence of affixes in French may make French readers more sensitive to this unit; and second, the depth of the English orthography, which makes GPC less reliable, may in turn favor the use of morphemes to increase the efficiency of English reading.

A key aspect of our study was to provide direct comparisons of how language and orthography impose variations in the use of morphemes in word recognition. This would contribute information about linguistic variation that would be useful in extending reading acquisition models to the morphological level. Our participants are typical readers in Grades 3 and 4, in other words, children who have already established early decoding in learning to read and who are expected to show morphemic effects on the basis of previous literature. However, we expect a degree of disparity between the groups due to cross-linguistic differences in relevant factors. The nature of these differences should help in understanding the impact of linguistic variation. We expect that orthographic depth and morphological productivity/transparency will both be influential. More use of morphemes would therefore be expected among the French group on the basis of morphological prevalence/ transparency but the question of whether the utility of morphemes in resolving the greater inconsistency in English will increase morphemic sensitivity in the English group beyond the level expected by the influence of morphological productivity/transparency has still to be resolved. In sum, if the presence of morphemes facilitates children's word recognition and if cognitive processing adapts to properties of environment stimuli, our first hypothesis is that children will rely on morphemic units when they process words and pseudowords, and our second hypothesis is that such morphological effects will be greater in the French language.

Method

Participants

Participants were 40 fourth graders from Scotland in the UK and 32 fourth graders from France. Both groups came from a similar middle income socioeconomic intake. The schools that we chose had middle-class catchment areas, according to national statistics in each country. Informed consent was obtained for each child. Mean age in the UK group was 8.41 years and mean age in the French group was 9.83 years. The difference was significant in terms of age (see **Table 1**) and not in terms of schooling because UK children start primary school 1 year before French children. A group of 32 French third graders matched for chronological age with the UK children was also recruited (see **Table 1**).

Background Measures

To ensure that the two groups of fourth graders were comparable in terms of language abilities, we assessed receptive vocabulary in each group using the *British Picture Vocabulary Scale* in the United Kingdom (Dunn et al., 1997) and the *Echelle de Vocabulaire en Images Peabody* in France (Dunn et al., 1993). All children performed within the normal range (percentiles 25–90). Reading skills were assessed using the *British Ability Scales Word Reading* subtest (Elliott et al., 1983) in the United Kingdom and the *Alouette Test* (Lefavrais, 1967) in France. All children performed within the normal range (percentiles 25–90). The UK group displayed a reading age greater than their chronological age (see **Table 1**).

Stimuli

A lexical decision task (LDT) was constructed following the same principles in both languages with close matching of stimuli for frequency, length and suffixes. We used the French Manulex database (Lété et al., 2004) and the English Children's Printed Word Database (CPWD, Masterson et al., 2003). There were four categories of words resulting from the presence or absence of a root and a suffix: (i) R+S+ [root and suffix, e.g., farmer (English), fermier [farmer] (French)]; (ii) R+S− [root but no suffix, e.g., window (English), boutique [shop] (French)]; (iii) R−S+ [no root but an (orthographic) suffix, e.g., murder (English), ménage [household] (French)]; and (iv) R−S− [no root and no suffix, e.g., narrow (English), pédale [pedal] (French)]. The items in condition (i) were the only real derivations. Pseudowords were formed from a similar principle resulting in four matched categories: (i) R+S+ [e.g., gifter (English), rosage (French)]; (ii) R+S− [e.g., puffow (English), lionque (French)]; (iii) R−S+ [e.g., gopter (English), mivage (French)]; and (iv) R−S− [e.g., ferbow (English), beadle (French)].

There were 29 items per condition in each language (see Appendix in Supplementary Material). Stimuli characteristics are presented in **Table 2**. There were 232 items in total. No fillers

TABLE 1 | Characteristics of the participants: mean chronological and reading age in years(range in brackets).

	English 4th gr	French 4th gr	French 3rd gr	En 4th gr-Fr 4th gr Student t	*p*-value	En 4th gr-Fr 3th gr Student t	*p*-value
N	40	32	32				
Chronological age	8.41 (7.58–9.25)	9.83 (9.33–10.58)	8.67 (7.58–9.25)	15.19	<0.001	0.95	0.21
Reading level	9.58 (6.5–14)	9.83 (8.5–11.83)	9.16 (7.67–10.91)	0.82	0.42	1.07	0.029

En, English; Fr, French; Gr, grade.

TABLE 2 | Stimuli characteristics in English and French languages.

WORDS	English	French	Difference student t	*P*-value
R+S+				
Frequency	46.83	41.46	0.27	ns
Length	6.38	7.10	2.99	0.01
R+: word frequency	106.14	156	1.16	ns
R+: length	4.21	5	4.3	0.01
R+S−				
Frequency	57.24	46.50	0.43	ns
Length	5.83	6.64	3.34	0.01
R+: word frequency	151.69	73.73	1.82	0.07
R+: length	3.72	3.89	0.75	ns
R−S+				
Frequency	59.97	45.49	0.81	ns
Length	6.241	6.59	1.33	ns
R−S−				
Frequency	44.55	44.38	0.04	Ns
Length	5.97	6.41	2.12	0.04
PSEUDOWORDS				
R+S+ length	6.35	7.04	2.55	0.01
R+S− length	5.83	6.79	4.41	0.00
R−S+ length	6.24	6.31	0.30	ns
R−S− length	5.97	6.10	0.61	ns

were added due to the length of the list. While this could potentially lead to an overestimation of the presence of embedded morphemes, our own assessment of the written language encountered by French children (via the Manulex database) indicates a high proportion of morpheme-like units. Due to differences in language characteristics, the roots in English derived words are nearly always complete, while roots are often truncated in French derived words. For example, the English word *farmer* contains the whole word root *farm*, while in the French word *fermier* [farmer], the final *e* of the root *ferme* has been removed. In our stimulus list, the whole lexical form of the root is truncated in 17 English words (10 in the R+S+ condition, 7 in the R+S− condition) and in 33 French words (22 in the R+S+ condition, 11 in the R+S- condition). In addition, in English, the base word is sometimes modified in the derived form by doubling the consonant. This is a peculiarity of English that complicates the orthographic definitions. However, given that this feature is quite common, it was also included as it was considered important to choose examples that were representative of the two languages in order to avoid concerns that our list of stimuli might be artificial in composition.

Procedure

The lexical decision task was administered using Cognitive Workshop software (Seymour, 1994–1999) in the UK and E-prime Software, Version 1.0 (Schneider et al., 2002). Items were presented centrally in lower case Courier New font, size 25. The participants were required to press the "YES" key (using their dominant hand) if the string was a real word, and a "NO" key (using the non-dominant hand) if the string was not a real word. A trial consisted of a fixation cross during 1500 ms and the target remained on the screen until the participant responded or for a maximum of 5000 ms. Reaction times were recorded via the keyboard. There were two counterbalanced sessions with 6 practice items. Items were presented in a randomized order for each participant. All items categories were mixed within one list. A short pause was introduced after every 20 items.

Results Part I: Grade-Level Matched Comparison

Data Analysis

Due to differences in the age of schooling, UK children in Grade 4 were a year younger than their French counterparts. Therefore, we decided to conduct the analyses in two parts. We first compared the performances of the UK group to those of the French children matched for grade, and then we compared the performances of the UK group to those of the French group matched for age.

Analyses of variance were performed on percentages of correct responses (accuracy) and reaction times to correct responses, with root (R+, R−) and suffix (S+, S−) as within-subjects factors and language group (UK, French) as between−subjects factors. Only responses longer than 400 ms and shorter than 5000 ms were considered in the analysis (0.2% of the data were discarded from the analysis). We conducted analyses by participants (F_1) and by items (F_2) and for the sake of clarity, only significant (or marginally non-significant) effects—at least on F_1 analyses—are reported.

Word Condition

Accuracy

Figure 1 presents the mean percentages of correct responses for word stimuli. French children performed more accurately than English children (95.61 and 75.97%, respectively), $F_{1(1,\ 71)} = 49.41$, $p < 0.001$, $\eta_p^2 = 0.41$, $F_{2(1,\ 224)} = 119.74$, $p < 0.001$, $\eta_p^2 = 0.35$. There was a main effect of suffix, $F_{1(1,\ 71)} = 23.00$, $p < 0.001$, $\eta_p^2 = 0.25$, $F_{2(1,\ 224)} = 3.77$, $p = 0.05$, $\eta_p^2 = 0.02$, and a root by suffix interaction in the analysis by participants only, $F_{1(1,\ 71)} = 3.35$, $p = 0.04$, $\eta_p^2 = 0.06$, $F_{2(1,\ 224)} = 2.50$, $p = 0.11$. As the root by suffix by language interaction was also significant (marginally so, by items), $F_{1(1,\ 71)} = 18.51$, $p < 0.001$, $\eta_p^2 = 0.21$, $F_{2(1,\ 224)} = 3.20$, $p = 0.07$, $\eta_p^2 = 0.02$, we examined this interaction in each group separately.

In the UK group, the root by suffix interaction was significant (marginally so, by items), $F_{1(1,\ 39)} = 18.19$, $p < 0.001$, $\eta_p^2 = 0.34$, $F_{2(1,\ 112)} = 2.88$, $p = 0.08$ $\eta_p^2 = 0.03$, indicating that while the presence of a suffix had no impact on accuracy when a root was present (R+S+: 76.03%, R+S−: 77.23%), it improved accuracy when there was no root (R−S+: 79.81%, R−S−: 70.81%). The effects were not significant in French.

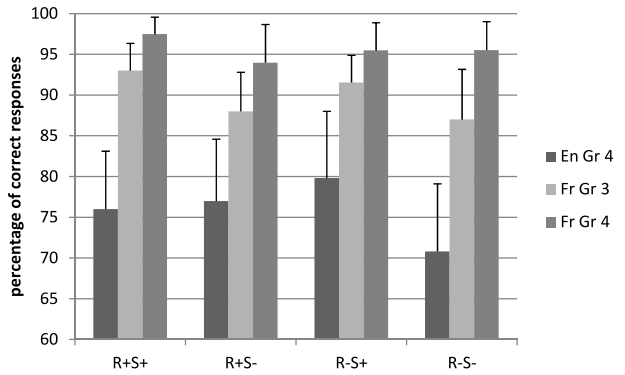


FIGURE 1 | Mean percentage of correct responses in the lexical decision task, real word conditions. En, English; Fr, French; Gr, grade; R+S+, Root present, suffix present; R+S−, Root present, suffix absent; R−S+, Root absent, suffix present; R−S−, Root absent, suffix absent.

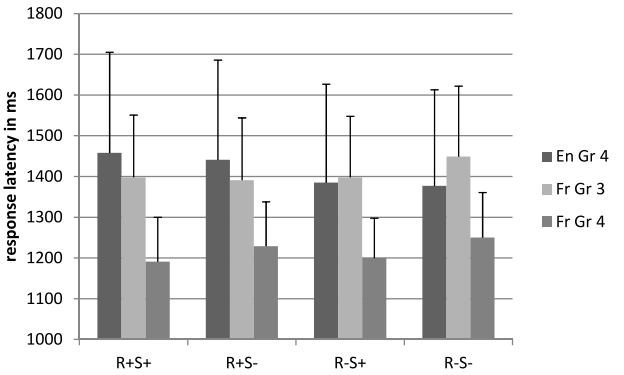


FIGURE 2 | Response latencies in the lexical decision task, real word conditions. En, English; Fr, French; Gr, grade; R+S+, Root present, suffix present; R+S−, Root present, suffix absent; R−S+, Root absent, suffix present; R−S−, Root absent, suffix absent.

Latencies

The mean latences for word stimuli.are reported in **Figure 2**. The French children responded faster than the English children [respectively, 1217 and 1415 ms, $F_{1(1,\ 71)} = 4.71$, $p = 0.03$, $\eta_p^2 = 0.06$, $F_{2(1,\ 224)} = 42.90$, $p < 0.001$, $\eta_p^2 = 0.16$]. There was a main effect of root [respectively, 1350 vs. 1317 ms, $F_{1(1,\ 71)} = 3.94$, $p = 0.05$, $\eta_p^2 = 0.05$ $F_{2(1,\ 224)} = 4.32$, $p = 0.04$, $\eta_p^2 = 0.02$. In addition, the root by language interaction was significant $F_{1(1,\ 71)} = 9.56$, $p = 0.003$, $\eta_p^2 = 0.05$, $F_{2(1,\ 224)} = 4.32$, $p = 0.04$, $\eta_p^2 = 0.02$]. Comparison showed a significant effect in English only, with the presence of a root slowing down latencies [R+: 1449 ms, R−: 1381 ms, $F_{1(1,\ 71)} = 12.32$, $p = 0.001$, $\eta_p^2 = 0.23$, $F_{2(1,\ 112)} = 6.28$, $p = 0.01$, $\eta_p^2 = 0.05$].

The suffix by language interaction was marginally significant by participants and non-significant by items, $F_{1(1,\ 71)} = 3.33$, $p = 0.07$, $\eta_p^2 = 0.05$, $F_{2(1,\ 224)} = 2.19$, $p = 0.14$, $\eta_p^2 = 0.01$. The presence of a suffix speeded up word recognition in French [1195 vs. 1239 ms, $F_{1(1,\ 29)} = 4.42$, $p = 0.04$, $\eta_p^2 = 0.13$, $F_{2(1,\ 112)} = 3.22$, $p = 0.07$, $\eta_p^2 = 0.03$] but not in English (1421 vs. 1409 ms), although it should be noted that this finding did not generalize across items.

Pseudoword Condition

Accuracy

The mean percentages of correct responses are displayed in **Figure 3**. French children responded more accurately than UK children [respectively, 98.8 and 64% correct, $F_{1(1,\ 71)} = 31.22$, $p < 0.001$, $\eta_p^2 = 0.31$, $F_{2(1,\ 224)} = 38.80$, $p < 0.001$, $\eta_p^2 = 0.28$]. There was a main effect of root, $F_{1(1,\ 71)} = 40.58$, $p < 0.001$, $\eta_p^2 = 0.36$, $F_{2(1,\ 224)} = 24.24$, $p < 0.001$, $\eta_p^2 = 0.10$, and an interaction between suffix and language, $F_{1(1,\ 71)} = 35.35$, $p < 0.001$, $\eta_p^2 = 0.33$, $F_{2(1,\ 224)} = 3.80$, $p = 0.05$, $\eta_p^2 = 0.02$. The root by suffix by language interaction was also significant by participants but not by items, $F_{1(1,\ 71)} = 39.12$, $p < 0.001$, $\eta_p^2 = 0.36$, $F_2 < 1$. For completeness, simple effects were used to investigate the interaction by participants further but it should be noted that the interaction did not generalize across items.

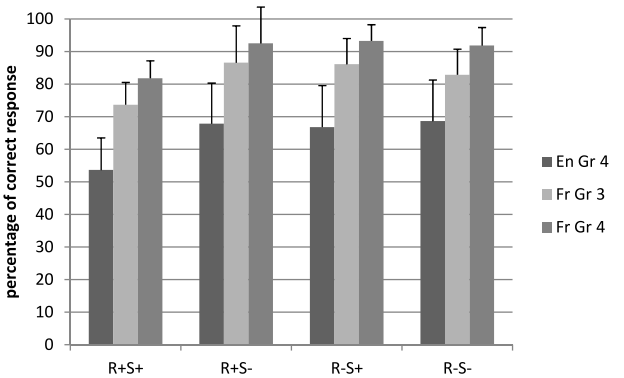


FIGURE 3 | Mean percentage of correct responses in the lexical decision task, pseudoword conditions. En, English; Fr, French; Gr, grade; R+S+, Root present, suffix present; R+S−, Root present, suffix absent; R−S+, Root absent, suffix present; R−S−, Root absent, suffix absent.

For the UK children, there were significant main effects of root, $F_{1(1,\ 42)} = 27.16$, $p < 0.001$, $\eta_p^2 = 0.39$, $F_{2(1,\ 112)} = 11.52$, $p < 0.001$, $\eta_p^2 = 0.09$, and suffix, $F_{1(1,\ 42)} = 25.80$, $p < 0.001$, $\eta_p^2 = 0.38$, $F_{2(1,\ 112)} = 19.60$, $p < 0.001$, $\eta_p^2 = 0.15$. There was also a root by suffix interaction, $F_{1(1,\ 42)} = 22.95$, $p < 0.001$, $\eta_p^2 = 0.35$, $F_{2(1,\ 112)} = 7.40$, $p = 0.008$, $\eta_p^2 = 0.06$, revealing that the effect of root was significant only when there was also a suffix present, with this combination of root plus suffix reducing pseudoword accuracy.

For French children, there were main effects of root, $F_{1(1,\ 39)} = 16.95$, $p < 0.001$, $\eta_p^2 = 0.37$, $F_{2(1,\ 112)} = 13.27$, $p < 0.001$, $\eta_p^2 = 0.11$ and suffix, $F_{1(1,\ 39)} = 15.82$, $p < 0.001$, $\eta_p^2 = 0.35$, $F_{2(1,\ 112)} = 9.46$, $p = 0.003$, $\eta_p^2 = 0.08$. The root by suffix interaction was significant, $F_{1(1,\ 39)} = 17.71$, $p < 0.001$, $\eta_p^2 = 0.38$, $F_{2(1,\ 112)} = 16.93$, $p < 0.001$, $\eta_p^2 = 0.13$, and, as for the UK group, the negative effect of the root only occurred when there was also a suffix present.

In all, the morphemic effects are similar in both languages but this interaction indicates that the effects (by participants) appear stronger in the UK children.

Latencies

Figure 4 shows the mean latencies for the pseudoword conditions. There was a main effect of root, $F_{1(1,\ 71)} = 6.37$, $p = 0.014$, $\eta_p^2 = 0.08$, $F_{2(1,\ 224)} = 13.96$, $p < 0.001$, $\eta_p^2 = 0.06$, and this effect interacted significantly with language, $F_{1(1,\ 71)} = 5.18$, $p = 0.03$, $\eta_p^2 = 0.07$, $F_{2(1,\ 224)} = 6.08$, $p = 0.01$, $\eta_p^2 = 0.04$. The negative impact of the root was present in French only [1727 vs. 1619 ms, $F_{1(1,\ 29)} = 9.34$, $p = 0.005$, $\eta_p^2 = 0.24$, $F_{2(1,\ 112)} = 16.60$, $p < 0.001$, $\eta_p^2 = 0.13$].

The suffix by language interaction was also significant by participants only, $F_{1(1,\ 71)} = 7.03$, $p = 0.01$, $\eta_p^2 = 0.09$, $F_{2(1,\ 224)} = 1.21$, $p = 0.27$, $\eta_p^2 = 0.02$. Across participants, this indicated that the presence of suffixes slowed down responses in the French group only [respectively, 1705 and 1681 ms, $F_{1(1,\ 29)} = 4.64$, $p = 0.04$, $\eta_p^2 = 0.14$, $F_{2(1,\ 112)} = 4.17$, $p = 0.04$, $\eta_p^2 = 0.04$].

Summary of Main Results, Part 1

In sum, sensitivity to morpheme units differed across languages in processing words, while patterns of response were more comparable for pseudoword processing.

Concerning words, the presence of a root slowed word recognition in English only. The presence of a suffix was only beneficial for English accuracy in the absence of a root. In French, the pattern was different: the presence of a suffix led to faster word recognition.

In pseudoword processing, across languages, reduced accuracy was observed when a pseudoword contained both a root and a suffix, although this effect was somewhat stronger in English. Only the French children showed latency effects, with the presence of either a root or a suffix leading to slower responses.

As the French children were younger than the UK children, a second analysis was conducted to match chronological age rather than school level.

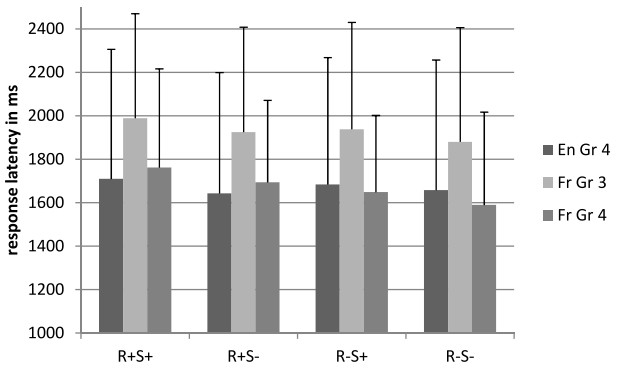


FIGURE 4 | Response latencies in the lexical decision task, pseudoword conditions. En, English; Fr, French; Gr, grade; R+S+, Root present, suffix present; R+S−, Root present, suffix absent; R−S+, Root absent, suffix present; R−S−, Root absent, suffix absent.

Results Part II: Chronological Age Matched Comparison

The results of the chronological age matched children are displayed in **Figures 1–4**.

Word Condition

Accuracy

The French children performed more accurately than the UK children [respectively, 89.99 vs. 75.97%, $F_{1(1,\ 71)} = 23.12$, $p < 0.001$, $\eta_p^2 = 0.78$, $F_{2(1,\ 224)} = 261.90$, $p < 0.001$, $\eta_p^2 = 0.70$], in spite of having received a year less of schooling. There was a main effect of suffix, $F_{1(1,\ 71)} = 34.60$, $p < 0.001$, $\eta_p^2 = 0.33$, $F_{2(1,\ 224)} = 6.07$, $p = 0.01$, $\eta_p^2 = 0.03$, a suffix by root interaction (marginal by items), $F_{1(1,\ 71)} = 8.01$, $p = 0.006$, $\eta_p^2 = 0.10$, $F_{2(1,\ 224)} = 2.85$, $p = 0.09$, and the root by suffix by language interaction was significant by participants only, $F_{1(1,\ 71)} = 11.37$, $p < 0.001$, $\eta_p^2 = 0.14$, $F_{2(1,\ 224)} = 1.48$, $p = 0.23$. For completeness, the interaction by participants was followed up using simple effects for each language group separately, although it should be noted that the interaction does not generalize across items.

The UK group showed a main effect of suffix in the analysis by participants, $F_{1(1,\ 41)} = 21.70$, $p < 0.001$, $\eta_p^2 = 0.34$, $F_{2(1,\ 112)} = 2.09$, $p = 0.15$ and an interaction between suffix and root (marginal by items), $F_{1(1,\ 41)} = 18.19$, $p < 0.001$, $\eta_p^2 = 0.30$, $F_{2(1,\ 112)} = 3.54$, $p = 0.07$, $\eta_p^2 = 0.07$, indicating that suffixes improved word recognition accuracy only when there was no root.

For the French group, only the main effect of suffix was significant, $F_{1(1,\ 29)} = 13.54$, $p < 0.001$, $\eta_p^2 = 0.32$, $F_{2(1,\ 112)} = 4.63$, $p = 0.03$, $\eta_p^2 = 0.04$: words with a suffix were recognized more accurately than words without suffix (92.20 vs. 87.50%, respectively).

Latencies

There was no main effect of language but the root by language interaction was significant, $F_{1(1,\ 71)} = 11.43$, $p = 0.001$, $\eta_p^2 = 0.14$, $F_{2(1,\ 224)} = 2.82$, $p = 0.03$, $\eta_p^2 = 0.14$: roots increased response latencies in the UK group only, $F_{1(1,\ 41)} = 12.32$, $p = 0.001$, $\eta_p^2 = 0.23$, $F_{2(1,\ 112)} = 6.28$, $p = 0.01$, $\eta_p^2 = 0.05$.

Pseudoword Condition

Accuracy

French children were more accurate than UK children [82.30 vs. 64% correct, respectively, $F_{1(1,\ 71)} = 14.17$, $p = 0.01$, $\eta_p^2 = 0.17$, $F_{2(1,\ 224)} = 8.96$, $p = 0.003$, $\eta_p^2 = 0.04$]. While there was no main suffix effect, the suffix by language interaction was significant by participants only, $F_{1(1,\ 71)} = 33.16$, $p < 0.001$, $\eta_p^2 = 0.32$. The root by suffix by language interaction was also significant only in the analysis by participants, $F_{1(1,\ 71)} = 52.12$, $p < 0.001$, $\eta_p^2 = 0.42$. Although this effect does not generalize across items, simple effects were used to understand the interaction by participants.

In the UK group, there were main effects of root, $F_{1(1,\ 42)} = 27.16$, $p < 0.001$, $\eta_p^2 = 0.39$, $F_{2(1,\ 112)} = 11.52$, $p < 0.001$, $\eta_p^2 = 0.09$, and suffix, $F_{1(1,\ 42)} = 25.80$, $p < 0.001$, $\eta_p^2 = 0.38$,

$F_{2(1,\ 112)} = 19.60$, $p < 0.001$, $\eta_p^2 = 0.15$, and an interaction between root and suffix, $F_{1(1,\ 42)} = 22.95$, $p < 0.001$, $\eta_p^2 = 0.35$, $F_{2(1,\ 112)} = 7.80$, $p = 0.008$, $\eta_p^2 = 0.06$, indicating that the combination of a root and a suffix decreased accuracy relative to other pseudowords.

In the French group, main effects of root, $F_{1(1,\ 29)} = 12.17$, $p = 0.002$, $\eta_p^2 = 0.30$, $F_{2(1,\ 112)} = 3.68$, $p = 0.05$, $\eta_p^2 = 0.04$, and suffix, $F_{1(1,\ 29)} = 11.64$, $p < 0.001$, $\eta_p^2 = 0.29$, $F_{2(1,\ 112)} = 4.85$, $p = 0.03$, $\eta_p^2 = 0.04$, were also observed, as well as an interaction between root and suffix, $F_{1(1,\ 29)} = 30.17$, $p < 0.001$, $\eta_p^2 = 0.51$, $F_{2(1,\ 112)} = 11.42$, $p < 0.001$, $\eta_p^2 = 0.09$. The interaction revealed reduced accuracy for the combination of a root and a suffix compared to other pseudowords.

This inspection of the data reveals that the suffix by root by group interaction (by participants) reflects the fact that the effects were stronger in French than in English.

Latencies

UK children responded faster than French children [respectively, 1674 and 1934 ms, $F_{1(1,\ 71)} = 6.77$, $p = 0.04$, $\eta_p^2 = 0.06$, $F_{2(1,\ 224)} = 146.08$, $p < 0.001$, $\eta_p^2 = 0.40$]. Across groups, response latencies were longer when a suffix was present [respectively, 1831 and 1777 ms, $F_{1(1,\ 71)} = 6.87$, $p = 0.011$, $\eta_p^2 = 0.09$, $F_{2(1,\ 224)} = 7.31$, $p = 0.007$, $\eta_p^2 = 0.03$].

Summary of Main Results, Part II

As in the comparison by grade level, there was indication of differential group sensitivity to morpheme units in word recognition but a similar pattern of pseudoword processing across languages.

For word recognition, only the UK children showed increased latencies when a root was present. Accuracy among the French children showed a higher degree of sensitivity to suffixes (regardless of whether a root was present or not).

For pseudoword processing, the effects were the same across languages: the presence of a suffix in a pseudoword slowed responses and the combination of a root plus a suffix reduced accuracy.

Discussion

Current models of reading development highlight cross-linguistic variation in naming accuracy in relation to early orthographic decoding (e.g., Ziegler and Goswami, 2005). However, these models do not offer an account of whether or not cross-linguistic effects operate on morphological processing during visual word recognition. The present study examined the extent to which morphemic effects in lexical access are universal or whether such effects can be modulated by language specificities during development.

For this purpose two comparable sets of lexical decision stimuli that manipulated the presence of component morphemes were presented to groups of French- and English-speaking children. As schooling starts 1 year earlier in the UK as compared to France, performance was first compared using a schooling match (Grades 4 in France and the UK), and in a second comparison, a chronological age match (Grade 3 in France and Grade 4 in the UK; both aged 8 years).

The data clearly indicate the importance of roots and suffixes for both language groups. Although the precise pattern differed, both groups were sensitive to the presence a suffix within words—either a genuine suffix or a suffix-like ending—which is consistent with the importance of suffixes as orthographic patterns. For pseudowords, the combination of a root with a suffix interfered with accurate processing in both languages. A tendency to slower responses was also observed when a pseudoword contained only a suffix, although this effect was clearer in French and present from Grade 3 onwards.

Cross-linguistic differences were also apparent, although some interaction effects were significant in the by-participants analysis only. In English, the presence of roots slowed down visual word recognition. Specific attention was given to the R+S+ vs. R+S- comparison, as these correspond respectively to the morphological and orthographic control conditions that are typically used in the literature on morphological decomposition in visual word recognition (see for example, Feldman et al., 2002; Casalis et al., 2009, for developmental studies). Interestingly, faster word recognition was observed when a suffix was present in the French analysis but not in the English analysis. This suggests a more specifically *morphological* sensitivity in French, whereas the results indicated sensitivity to embedded words in English, since roots were mostly free-standing words.

In English, suffixes only affected the accuracy of word recognition in the absence of a root; whereas, in French, suffixes generally led to faster word recognition and, for the older Grade 4 group, improved accuracy only when combined with a root (i.e., the R+S+ real derivations, e.g., farmer). This latter effect of school grade in French suggests that reading skills and/or language proficiency has an impact on suffix processing.

A detrimental effect of the root was observed in English only. This effect was not apparent in French as the impact of the root produced only facilitation effects among French children. This cross-language discrepancy may derive from the fact that, in most cases, roots corresponded to whole words in English (41 out 58 items), whereas this was less true of French (25 out 58 items). This would be consistent with Nation and Cocksey's (2009) finding of an automatic semantic activation of embedded words among English-speaking 7-year-olds. Therefore, the inhibition effect observed in the present study may reflect processing costs associated with identification of the root and competition with whole word processing. Indeed, a striking finding is that the inhibition effect in English is observed in both R+S− words, which may be considered to be orthographic control items, and R+S+, which are derived forms. Morphological priming studies report only facilitation effects, both among skilled readers of English (e.g., Rastle and Davis, 2008) and developing readers of French (Quémart et al., 2011). Minimally, then the inhibition effects observed here indicate that young English-speaking readers are sensitive to embedded word units in visual word recognition. While higher frequency embedded words in English might have favored an inhibition effect in English, the languages did not differ significantly in this respect in either the R+S+ or R+S = conditions although it should be noted that the outcome was

marginal ($p = 0.07$) in the R+S− condition. While French and English stimuli were statistically matched in terms of frequency, French words tended to be slightly longer than English words. Although the difference was less than 1 letter on average, this could potentially have led French children to show more reliance on a decomposition strategy for word processing. Another source of difference lies in the fact that some suffixes have been repeated, leading potentially to an increased sensitivity to morphological decomposition. Note that slightly more repetition occurred in English (-er) than French.

Strong effects of morpheme units were found in pseudoword processing where the combination of both root and suffix led to an increasing rate of errors. This result is in line with previous research, suggesting that young readers rely on morpheme units when they have to process an unknown word (Burani et al., 2002; Quémart et al., 2012). At the same time, linguistic variation also came into play as the beneficial effects of suffixes, in particular, were stronger in French pseudoword processing.

A methodological difficulty when comparing children from different countries is that such a comparison goes beyond differences in native language. A first issue is that schooling starts during the fifth year in UK while it starts during the sixth year in France. This was dealt with by performing two separates analyses: one based on a school-level matched design, with French children being older than UK children; and the other based on a chronological-age matched design, with the UK children having experienced a year more of schooling. It was not possible to achieve a perfect matching between the groups as the English-speaking children were less accurate than the French children regardless of the method of matching groups. In contrast, the UK children exhibited slower latencies in word processing than the older Grade 4 French children in the first analysis but were faster at pseudoword processing than the Grade 3 French children in the second analysis. A second issue is connected to the school curriculum, particularly in relation to the teaching of reading and morphology. Across languages, our participants all came from schools adopting a mixed method approach to reading instruction: whole-word and phonics. In France, morphological structure is explicitly taught at Grade 4 and, in the Scottish education system that the UK children experienced, intensive instruction about derivational affixes begins in Grades 3 or 4 as part of spelling instruction. Further studies should address instructional issues in a more systematic way. Our study was a first attempt to directly compare the use of morphological units across languages. It will therefore be necessary to extend this work to larger samples as well as other languages.

In terms of group differences in word processing, French children were always more accurate, and were faster only when they were older (schooling matching); in pseudoword processing, French children again always responded more accurately, but responded more slowly when they were matched on age (with less schooling). Note that the difference may be explained by the fact that the French pseudowords were almost one letter longer than the English pseudowords. However, it is possible that the additional year of schooling experienced by the UK children may also have contributed to their faster pseudoword reaction times. Beyond these group differences, both analyses yielded quite similar patterns of results. However, the slight differences that emerged between Part 1 and Part 2 reveal that morphological processing develops during schooling. More specifically, the presence of a root slowed down latencies for fourth graders only (UK), and there were more indication of a suffix benefit among French fourth graders than French third graders.

Thus, our study demonstrates that developing readers make use of morphology when recognizing familiar words and when processing new words. In a previous study, Duncan et al. (2009) compared English and French morphological awareness in relation to derivation with suffixes. The results clearly showed that the UK children were outperformed by the French children when they had to manipulate morpheme units explicitly. Note that sensitivity to morphemes, as assessed by a relational judgment task, was found to be similar in both groups. These results were interpreted with reference to the importance of morphological structure in French. It is therefore interesting to note that, in the present study, UK children make use of morphemes during lexical access, even though overall they were less accurate at this than French children and were less sensitive to true derivations (R+S+ words). This outcome aligns with two conclusions: first, morphemes may be used in word and pseudoword processing regardless of GPC transparency; and second, when confronted with a rich morphological system, children may develop morphological knowledge faster and acquire a more finely-tuned sensitivity to written morphology.

In conclusion, research on reading acquisition reflects a growing interest in morphological processing, as once children have completed the first phases of reading acquisition they are confronted by a growing number of long and derived words. Previous research on phonological coding in reading aloud has pointed to the importance of cross-language variation in the nature and speed of acquisition of GPC, and was formalized in the PGST. Our intention was to begin the process of examining morphological processing in visual word recognition within a similar framework. The languages under investigation differed in terms of orthographic depth, with English being more opaque than French, and morphological productivity, with French being morphologically richer than English. The first aim was to examine whether morphology was generally used by developing readers in Grades 3 and 4. A main result was that children make use of morphemic information in both languages confirming our first hypothesis of the relevance of morphology in reading development in both opaque and more transparent alphabetic orthographies. The second aim was to assess the importance of two factors expected to be influential, namely, orthographic depth and morphological prevalence/transparency. Both aspects could be contrasted in English and French in opposing directions, with the English orthography being more opaque and French morphology being richer. One key question was therefore whether the utility of morphemes in resolving the greater inconsistency in English increased sensitivity in this group beyond the level expected by the influence of morphological productivity/transparency. Results indicated stronger morphological effects in French, confirming our hypothesis that morphological richness will

outweigh orthographic depth at least in alphabetic writing systems.

Acknowledgments

This research was completed while the first author was a doctoral student, funded by the French Ministry of Research and Technology, and the third author was a visiting lecturer, both at the University of Lille, France. This project was supported by the grant LECT MORPHO from the National Agency for Research (ANR) award to SC and the grant from the French Research Agency ANR−11-EQPX-0023, FEDER SCV-IRDIVE and University Lille 3. We also thank Marion Janiot and Elaine Gray for help with participant recruitment and testing.

Supplementary Material

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2015.00452/abstract

References

Bauer, L. (2003). *Introducing Linguistic Morphology, 2nd Edn*. Edinbourg: Edinburgh University Press.

Burani, C., Marcolini, S., and Stella, G. (2002). How early does morpholexical reading develop in readers of a shallow orthography? *Brain Lang.* 81, 568–586. doi: 10.1006/brln.2001.2548

Burani, C., Marcolini, S., De Luca, M., and Zoccolotti, P. (2008). Morpheme-based reading aloud: evidence from dyslexic and skilled Italian readers. *Cognition* 108, 243–262. doi: 10.1016/j.cognition.2007.12.010

Carlisle, J. F., and Katz, L. (2006). Effects of word and morpheme familiarity on reading of derived words. *Read. Writ.* 19, 669–693. doi: 10.1007/s11145-005-5766-2

Carlisle, J. F., and Stone, C. A. (2005). Exploring the role of morphemes in word reading. *Read. Res. Q.* 40, 428–449. doi: 10.1598/RRQ.40.4.3

Casalis, S., Dusautoir, M., Colé, P., and Ducrot, S. (2009). Morphological relationship to children word reading: a priming study in fourth graders. *Br. J. Dev. Psychol.* 27, 761–766. doi: 10.1348/026151008X389575

Clark, E. V. (1998). Lexical creativity in French-speaking children. *Cahiers de Psychologie Cognitive* 17, 513–30.

Colé, P., Bouton, S., Leuwers, C., Casalis, S., and Sprenger-Charolles, L. (2012). Stem and derivational-suffix processing during reading by French second and third graders. *Appl. Psycholinguist.* 33, 97–120. doi: 10.1017/S0142716411000282

Coltheart, M., and Crain, S. (2012). Are there universals of reading? We don't believe so. *Behav. Brain Sci.* 35, 282–283. doi: 10.1017/S0140525X12000155

Deacon, S. H. (2012). Bringing development into a universal model of reading. *Behav. Brain Sci.* 35, 284–284. doi: 10.1017/S0140525X12000040

Deacon, S. H., Whalen, R., and Kirby, J. R. (2011). Do children see the danger in dangerous? Grade 4, 6, and 8 children's reading of morphologically complex words. *Appl. Psycholinguist.* 32, 467–481. doi: 10.1017/S0142716411000166

Duncan, L. G., Casalis, S., and Colé, P. (2009). Early meta-linguistic awareness of derivational morphology: some observations from a comparison of English and French. *Appl. Psycholinguist.* 30, 405–440. doi: 10.1017/S0142716409090213

Duncan, L. G., Castro, S. L., Defior, S., Seymour, P. H. K., Baillie, S., Leybaert, J., et al. (2013). Phonological development in relation to native language and literacy. variations on a theme in six alphabetic orthographies. *Cognition* 127, 398–419.

Dunn, L. M., Dunn, L. M., Whetton, C., and Burley, J. (1997). *British Picture Vocabulary Scale II*. Windsor: NFER-Nelson.

Dunn, L. M., Theriault-Whalen, C. M., and Dunn, L. M. (1993). *Echelle de Vocabulaire en Images Peabody, Adaptation Française*. Toronto, ON: Psycan.

Elliott, C. D., Murray, D. J., and Pearson, L. S. (1983). *British Ability Scales*. Windsor: NFER Nelson.

Feldman, L. B., Rueckl, J., DiLiberto, K., Pastizzo, M. J., and Vellutino, F. R. (2002). Morphological analysis by child readers as revealed by the fragment completion task. *Psychon. Bull. Rev.* 9, 529–535. doi: 10.3758/BF03196309

Frost, R. (2012). Towards a universal model of reading. *Behav. Brain Sci.* 35, 263–279. doi: 10.1017/S0140525X11001841

Frost, R., Katz, L., and Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: a multilingual comparison. *J. Exp. Psychol. Hum. Percept. Perform.* 13, 104–115. doi: 10.1037/0096-1523.13.1.104

Goswami, U., and Bryant, P. (1990). *Phonological Skills and Learning to Read*. London: Erlbaum.

Hulme, C., and Snowling, M. J. (2013). The interface between spoken and written language: developmental disorders. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 369:20120395. doi: 10.1098/rstb.2012.0395

Laxon, V., Rickard, M., and Coltheart, V. (1992). Children read affixed words and non-words. *Br. J. Psychol.* 83, 407. doi: 10.1111/j.2044-8295.1992.tb02450.x

Lefavrais, P. (1967). *Test de l'Alouette (the Alouette test)*. Paris: Editions du Centre de Psychologie Appliquée.

Lété, B., Sprenger-Charolles, L., and Colé, P. (2004). MANULEX: a grade-level lexical database from French elementary school readers. *Behav. Res. Methods Instr. Comput.* 36, 156–166. doi: 10.3758/BF03195560

Mann, V., and Singson, M. (2003). "Linking morphological knowledge to English decoding ability: large effects of little suffixes" in *Reading Complex Words: Cross-Languages Studies*, eds E. M. H. Assink and D. Sandra (New York, NY: Kluwer Academic), 1–26.

Marcolini, S., Traficante, D., Zoccolotti, P., and Burani, C. (2011). Word frequency modulates morpheme-based reading in poor and skilled Italian readers. *Appl. Psycholinguist.* 32, 513–532. doi: 10.1017/S0142716411000191

Masterson, J., Stuart, M., Dixon, M., and Lovejoy, S. (2003). *The Children's Printed Word Database*. Available online at: www.essex.ac.uk/psychology/cpwd

Melby-Lervåg, M., Lyster, S. A., and Hulme, C. (2012). Phonological skills and their role in learning to read: a meta-analytic review. *Psychol. Bull.* 138, 322–352. doi: 10.1037/a0026744

Muter, V., Hulme, C., Snowling, M. J., and Stevenson, J. (2004). Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development: evidence from a longitudinal study. *Dev. Psychol.* 40, 665–681. doi: 10.1037/0012-1649.40.5.665

Nagy, W. E., and Anderson, R. C. (1984). How many words are there in printed school in English? *Read. Res. Q.* 19, 304–330. doi: 10.2307/747823

Nation, K., and Cocksey, J. (2009). Beginning readers activate semantics from sub-word orthography. *Cognition* 110, 273–278. doi: 10.1016/j.cognition.2008.11.004

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

Quémart, P., Casalis, S., and Duncan, L. G. (2012). Exploring the role of bases and suffixes when reading familiar and unfamiliar words: evidence from French young readers. *Sci. Stud. Read.* 16, 424–442. doi: 10.1080/10888438.2011.584333

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

Rey-Debove, J. (1984). Le domaine de la morphologie lexicale [The domain of lexical morphology]. *Cahiers de Lexicologie* 45, 3–19.

Schneider, W., Eschmann, A., and Zuccolotto, A. (2002). *E-Prime Reference Guide*. Pittsburgh, PA: Psychology Software Tools.

Seymour, P. H. K. (1994–1999). *Cognitive Workshop Software*. Developed at the School of Psychology, University of Dundee by Software Interventions Ltd. in cooperation with The Child Research Centre at the University of Jyväskylä, Finland.

Seymour, P. H. K., Aro, M., and Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *Br. J. Psychol.* 94, 143–174. doi: 10.1348/000712603321661859

Ziegler, J. C., and Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory. *Psychol. Bull.* 131, 3–29. doi: 10.1037/0033-2909.131.1.3

Ziegler, J. C., and Goswami, U. (2006). Becoming literate in different languages: similar problems, different solutions. *Dev. Sci.* 9, 429–436. doi: 10.1111/j.1467-7687.2006.00509.x

Ziegler, J. C., Perry, C., Jacobs, A. M., and Braun, M. (2001). Identical words are read differently in different languages. *Psychol. Sci.* 12, 379–384. doi: 10.1111/1467-9280.00370

Ziegler, J. C., Stone, G. O., and Jacobs, A. M. (1997). What's the pronunciation for _OUGH and the spelling for /u/? A database for computing feedforward and feedback inconsistency in English. *Behav. Res. Methods Instr. Comput.* 29, 600–618. doi: 10.3758/BF03210615

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 ARTICLE
published: 19 August 2014
doi: 10.3389/fpsyg.2014.00903

Does the mean adequately represent reading performance? Evidence from a cross-linguistic study

Chiara V. Marinelli[1], Joanna K. Horne[2], Sarah P. McGeown[3], Pierluigi Zoccolotti[4,5] and Marialuisa Martelli[1,4]*

[1] IRCCS Fondazione Santa Lucia, Rome, Italy
[2] Department of Psychology, University of Hull, Hull, UK
[3] School of Education, Edinburgh University, Edinburgh, UK
[4] Department of Psychology, Sapienza University of Rome, Rome, Italy
[5] Institute of Cognitive Sciences and Technologies, ISTC-CNR, Rome, Italy

Edited by:
Davide Crepaldi, University of Milano-Bicocca, Italy

Reviewed by:
Alessio Toraldo, Pavia University, Italy
Serje Robidoux, Macquarie University, Australia

****Correspondence:***
Marialuisa Martelli, Department of Psychology, Sapienza University of Rome, Via dei Marsi 78, 00176 Rome, Italy
e-mail: marialuisa.martellii@uniroma1.it

Reading models are largely based on the interpretation of average data from normal or impaired readers, mainly drawn from English-speaking individuals. In the present study we evaluated the possible contribution of orthographic consistency in generating individual differences in reading behavior. We compared the reading performance of young adults speaking English (one of the most irregular orthographies) and Italian (a very regular orthography). In the 1st experiment we presented 22 English and 30 Italian readers with 5-letter words using the Rapid Serial Visual Presentation (RSVP) paradigm. In a 2nd experiment, we evaluated a new group of 26 English and 32 Italian proficient readers through the RSVP procedure and lists matched in the two languages for both number of phonemes and letters. The results of the two experiments indicate that English participants read at a similar rate but with much greater individual differences than the Italian participants. In a 3rd experiment, we extended these results to a vocal reaction time (vRT) task, examining the effect of word frequency. An ex-Gaussian distribution analysis revealed differences between languages in the size of the exponential parameter (tau) and in the variance (sigma), but not the mean, of the Gaussian component. Notably, English readers were more variable for both tau and sigma than Italian readers. The pattern of performance in English individuals runs counter to models of performance in timed tasks (Faust et al., 1999; Myerson et al., 2003) which envisage a general relationship between mean performance and variability; indeed, this relationship does not hold in the case of the English participants. The present data highlight the importance of developing reading models that not only capture mean level performance, but also variability across individuals, especially in order to account for cross-linguistic differences in reading behavior.

Keywords: reading, individual differences, cross-linguistic comparison

INTRODUCTION

Reading is a complex task that involves several cognitive and sensory-motor components from image detection to the comprehension of meaning. It takes years to master this skill and during this progression, each of the components undergoes maturation and specific learning effects. Literate adults read with near perfect accuracy at an impressive speed, optimizing each of the processes involved and performing them in parallel. The speeding up of the function may be seen as moving from serial to parallel analysis up to the point in which individuals learn to master orthographic decoding of a letter string in a glance (e.g., Ziegler and Goswami, 2005).

In 1992, Carver proposed a bold conjecture to account for reading rate. Carver showed that readers adjust their reading rate, speeding up if they are searching for a particular word in a text (scanning) and slowing down if they want to memorize concepts. According to Carver, readers may shift "gear" to achieve the desired goal, but they generally read in the middle (third) gear or "rauding" (i.e., reading and auding) which optimizes comprehension considering the speed limits set by the processing components. In a classic paper, Taylor (1965) surveyed the reading skills of 12,000 US students, from first grade to college, and found the average rate to be 300 words per minute (wpm), which was taken by Carver (1992) as an estimate of rauding rate.

This functional measure of reading speed incorporates several components from decoding to motor execution, and it is relatively stable across individuals. Notably, it has been shown that pronunciation time, the most time consuming process, weights heavily on the average speed but contributes minimally to individual differences (Martelli et al., 2014). This means that pronunciation time adds a substantial constant factor to the (much faster) compartment of decoding. Furthermore, it indicates that the maximal

reading rate obtained in standard conditions (i.e., rauding) does not necessarily indicate the maximum processing rate for each of the sub-components in reading. Put in different terms, the articulatory component (as well as the eye movement scanning; see below) may pose an upper bound to the estimate of maximal reading rate that can be obtained in functional reading.

In a different line of research, focussed on assessing the perceptual limitations in reading, several authors measured reading speed by means of the Rapid Serial Visual Presentation (RSVP) paradigm. In this procedure, a sequence of words is rapidly presented in the same retinal location. The observer is required to name the words presented (typically a stream of four words per trial) without a time limit. The duration of the words on the screen to achieve a certain level of task performance (typically 80%) is measured. In this paradigm, the articulatory components do not directly exert a role on the estimation of the reading rate, since no time limit is given to complete the response. Furthermore, unlike ordinary reading, the observer does not have to scan for the words to read by eye movements, as stimuli are all presented in the same retinal position. Thus, this procedure minimizes the role of memory, pronunciation time and eye movements, allowing a more direct examination of the decoding components in reading (see Rubin and Turano, 1992; Chung et al., 1998; Legge et al., 2001; Pelli et al., 2007). In fact, compared to other reading techniques, RSVP gives the opportunity to substantially "speed up" reading rate. For example, Potter (1984) originally showed that reading and recall is still excellent at 12 words per second (i.e., 720 wpm), which is much faster than the level of "rauding."

In the absence of specific reading or visual deficits, and controlling the stimuli for high level cognitive factors, one may assume that decoding is similar across individuals. Indeed, most low-level visual functions, such as acuity or contrast sensitivity, are similar across subjects (Barlow, 1962; Fisher, 1975; Pelli et al., 2006; Strasburger et al., 2011), revealing that perceptual limitations are invariant across individuals and labs. However, when considering the reading speed measurements obtained with RSVP, variability across subjects and labs is, surprisingly, very large. In some cases, the advantage given by the RSVP technique in speeding up reading rate is relatively low, with reading rates around 300 wpm (Latham and Whitaker, 1996; Fine et al., 1997, 1999; Chung et al., 1998; Pelli et al., 2007) while, in other studies, reading rates exceeding 1500 wpm have been reported (Rubin and Turano, 1992; Latham and Whitaker, 1996).

Part of the large discrepancy in RSVP reading across labs may be related to low-level visual effects, such as presence/absence of masking (Felsten and Wasserman, 1980; Breitmeyer, 1984; Enns and Di Lollo, 2000) or to the number of items used in the stream. In particular, in some studies, four words are presented per trial, while in others, number of words well exceeds the memory span (e.g., Latham and Whitaker, 1996; Chung et al., 1998; Yager et al., 1998; Fine et al., 1999; Kwon et al., 2007; Pelli and Tillman, 2007; Pelli et al., 2007; Yu et al., 2007, 2010; Lee et al., 2010; Kwon and Legge, 2012). Note that these studies are mainly concerned with factors affecting visual limitations to reading, such as font size or letter spacing, and much less to cognitive dimensions (as well as to absolute estimates of reading rate which are rarely commented on). Thus, direct comparisons between the various estimates are hard to make since the stimuli are usually not designed to take into consideration linguistic variables (e.g., word frequency, orthographic complexity, orthographic neighborhood, age of acquisition, etc.) that are known to influence speed of reading (e.g., Coltheart et al., 1977; Ferrand and New, 2003).

Furthermore, there is also a surprisingly large discrepancy in reading rate across languages, such as when comparing the irregular English orthography with the consistent Italian one with similar RSVP reading tasks. The reading rate of English 5th and 7th graders with the RSVP of stimuli averages at around 500 wpm (Kwon et al., 2007), while normal 6th grade Italian readers do not exceed 120 wpm, a rate much slower than any other reported for this age level (Martelli et al., 2009). Italian dyslexics' average reading rate is as slow as 40 wpm (Martelli et al., 2009). Although suggestive, comparisons between these two languages are certainly difficult to interpret across experiments, particularly since Italian words tend to be long and morphologically complex, while English words tend to be shorter and morphologically simple.

As described above, most studies on reading focus on group data that average across participants and trials, and only recently it has been suggested that "*it is possible that some of the inconsistencies in the literature may be driven by individual differences among participants*" (Yap et al., 2012, p. 2). The source of this variability may possibly concern strategic differences related to the linguistic demands (both within a language and across different languages). Following Yap et al. (2012), we conjecture that, over and above differences in average speed, variability estimates may also provide insights into the computation involved in reading. Here, we were interested in exploring such variability in relation to differences in orthographic consistency, with the ultimate goal of understanding the invariant and variable properties of reading across languages. Indeed, learning to become a proficient reader in different orthographies may pose very different requirements to the reader and the end product of these different task demands may well be expressed by different degrees of inter-individual variability.

In the present study, we address a number of questions, comparing Italian and English readers. Is there a difference in processing speed of regular and irregular orthographies, once most of the cognitive variables are taken into account? Does the general speed factor interact with the efficiency of the orthographic decoding, as reflected by the size of the lexical effects in the two languages? Indeed, Faust et al. (1999) showed that larger effects of the experimental manipulations are expected in the case of differences in overall processing time across individuals (i.e., larger effects for slower individuals). Do the individual differences across languages arise from different strategies adopted in reading? The difference engine model (DEM), proposed by Myerson et al. (2003), explains group RT differences by assuming that, in the absence of a peripheral deficit, most differences between individuals are due to the amount of cognitive processing required predicting the relationship between mean and SD. Is this relationship as well as vRTs distribution similar across languages in the case of reading tasks? In this study, we attempt to answer these questions through three experiments that compared reading

speed (assessed with either the RSVP procedure or with vRT measurements) in a very regular (Italian) and in a very irregular (English) orthography with controlled orthographic materials.

EXPERIMENT 1: PROCESSING SPEED DIFFERENCES BETWEEN ENGLISH AND ITALIAN READERS

In this first preliminary experiment, we aim to explore possible differences in processing speed between Italian and English proficient readers, controlling for as many psycholinguistic variables as possible, based on the structural differences between the two languages. Previous observations report large discrepancies in RSVP reading rate across labs and languages, with English observers obtaining much higher estimates of reading rate (e.g., Rubin and Turano, 1992; Latham and Whitaker, 1996; Chung et al., 1998; Kwon et al., 2007; Martelli et al., 2009). However, due to concurrent procedural differences and uncontrolled variables, it is hard to draw a firm conclusion on these data. Here, we test a group of English young adults and a group of peer Italian readers using the RSVP paradigm to confirm the possible presence of different reading rates.

MATERIALS AND METHODS

Participants

Thirty Italian (15 males and 15 females) and 22 English (11 males and 11 females) readers participated in this experiment. Participants were university students recruited from the student population of the Sapienza University of Rome in Italy and of the University of Hull in the United Kingdom. Groups were comparable for age and gender. The age of the Italian group ranged between 19 and 28 years (mean age: 22.96, $SD = 2.84$) with 15.81 ($SD = 1.39$) years of schooling; the age of the English participants ranged between 18 and 24 years (mean age: 20.86, $SD = 3.77$) with 14.23 ($SD = 1.02$) schooling years. All participants were self-reported good readers, without a history of language, reading or spelling disorders. This study, as well as the ones presented in Experiments 2 and 3 (both conducted according to the principles of the Helsinki Declaration) were approved by the Ethical Committee of the Department of Psychology of Rome, and by that of the University of Hull in line with the BPS guidelines. Before taking part in the experiment, the subjects were given a description of the study and approved their participation.

Stimuli, apparatus, and procedure

In both languages, words were all nouns, without morphological complexity and irregularity in grapheme-to-phoneme correspondence. Because stress assignment to Italian polysyllabic words is unpredictable by rule, no words with irregular stress were included in the Italian list (i.e., all words were stressed on the syllable before the last). No irregular stress words were used also in the English list. In both languages, archaic, obsolete, poetic and scientific forms were avoided. For the Italian readers, a list of 80 5-letter words were selected from the LEXVAR database (Barca et al., 2002) with frequency ranging from 0 to 100 (mean frequency = 25.1, $SD = 24.4$, Colfis database; Bertinetto et al., 2005). For the English readers, a list of 80 5-letter words was selected from the MRC Psycholinguistic Database 2.0 (Wilson, 1988): Frequency ranged from 0 to 100 (mean frequency: 24.8, $SD = 36.2$ CELEX database, Baayen et al., 1993). Note that, to compare the frequency values of the two databases (the English database has one million of occurrences, while the Italian database counts over three million occurrences), the Italian word frequency values were reported to one million of occurrences. In Appendix A, means (and SDs) of the psycholinguistic variables are reported for the Italian and English lists. The Italian and English lists were matched for frequency, n-size, imageability and age of acquisition (all $ps > 0.1$). Italian and English words were comparable for bigram frequency based on values reported in the MCWord database (Medler and Binder, 2005) for English and in the LEXVAR database (Barca et al., 2002) for Italian language (referring to one million of occurrences). As it can be seen in Appendix A, lists were not matched between languages for number of phonemes [$t_{(159)} = 7.92$, $p < 0.0001$], that was higher for the Italian than the English list. Moreover, it was not possible to match the lists for number of syllables [$t_{(159)} = 14.41$, $p < 0.0001$], as English and Italian differ in the number of syllables and in the complexity of the syllabic structure. The number of syllables is generally higher in Italian (the mode length in the Italian lexicon, according to De Mauro, 1999, is 4 syllables) than in English. Moreover, in English, only 5% of monosyllables are CV (De Cara and Goswami, 2002), while in Italian (as in other romance languages) CV is the most frequent syllable type, covering 56% of syllable tokens in written corpora (for a more detailed description of Italian see Burani et al., 2014; for English, see Wyse and Goswami, 2008).

Words were rendered in Courier New font, a proportionally spaced font, and each letter subtended 0.4° of visual angle. Participants were seated 57 cm away from the computer screen (refresh rate = 60 Hz). A fixation point (a black square subtending 0.2° of visual angle) was presented at the center of the screen for 2000 ms. Immediately after the offset of fixation point, words were presented using the RSVP paradigm, i.e., four words were presented sequentially, one word at a time, at the same location on the display and participants were asked to read them aloud. There was no blank frame (zero inter-stimulus interval) between words. Following Rubin and Turano (1992) no mask was presented prior to the first or after the fourth word in the stream. We measured the duration threshold for each participant by varying exposure duration in a 20-trial run using the improved QUEST staircase procedure with a threshold criterion of 80% correct responses (Watson and Pelli, 1983). The adaptive QUEST procedure increased or decreased the presentation rate (starting from 500 ms) according to the participant's accuracy. Word omissions, mispronunciations and substitutions were considered to be errors. In order for the subjects to familiarize with the RSVP paradigm 10 practice trials (40 4-letter words) were administered prior to the beginning of the experiment. As in the experimental session the word duration in each trial was controlled by the adaptive procedures based on response accuracy.

RESULTS

The reading rate (i.e., wpm) was measured as 60/duration threshold∗1000 using the geometric mean as measure of the central tendency of the distribution (represented using a log scale, **Figure 1** lower axis) and the 95% confidence intervals (CIs) to

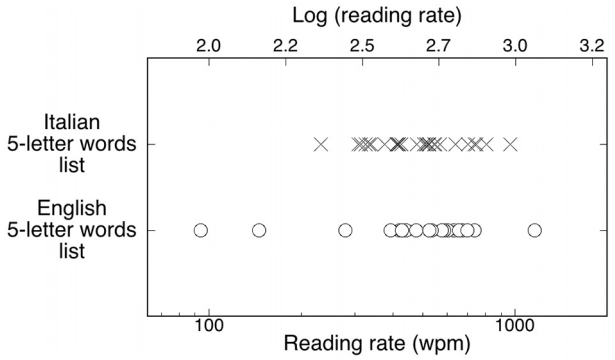


FIGURE 1 | Individual reading rates for Italian readers (Xs) and English participants (open circles) for a letter-matched list of words. The upper scale shows the corresponding values expressed as a log (wpm). Note the greater dispersion of experimental points among English than Italian observers.

express the variability in the distributions. ANOVA comparisons across groups were performed on log-transformed reading rates (linear scale, **Figure 1** upper axis). The reading speed for the English list (geomean = 449 wpm; *CI*: 346–583) was not different from the reading speed of the Italian parallel list [479 wpm; *CI*: 433–548; $F_{(1,50)} < 1$, $p = 0.55$]. Results were replicated also when socio-demographic variables (i.e., gender, age, and years of schooling) were added to the analysis as covariates: the main effect of the language factor was not significant [$F_{(1,44)}$ about 1; $p = 0.31$]; furthermore, none of the covariates were significant.

Figure 1 presents the reading rate distributions for the Italian and English readers. An inspection of the figure indicates that the English group was less homogeneous than the Italian group, with a larger variability: the group comprised the fastest individual and individuals who were slower (by a factor of ca. two) than the slowest Italian reader. This pattern is confirmed by the Levene's test for equality of variances: the variances of the Italian and English samples were significantly different ($F = 4.17$, $p < 0.05$).

As variability appears as the key feature of the group differences between the two languages we replicated this analysis using untransformed threshold values to check whether the difference in the variance of the two distributions could be due to the adoption of a nonlinear transformation. Mean duration thresholds were 120 ms ($SD = 44$) for the Italian group and 152 ms ($SD = 142$) for the English group. Again, the variances of the two groups were significantly different ($F = 8.86$, $p < 0.01$). Therefore, it appears that the difference in variability between the two languages is not due to the use of a nonlinear transformation of data.

Comments

Contrary to expectations based on our preliminary observations, and the work of Paulesu et al. (2000), Italian readers as a group were neither faster nor slower than English readers once the items were made comparable for some relevant psycholinguistic variables. However, the two groups showed substantial differences in individual variability with the Italian group considerably more homogeneous than the English one. The English group included both the fastest participant, reading over 1100 wpm, and the slowest participant, reading at ca. 90 wpm. Clearly, this phenomenon is captured by the variability in the two distributions and not by the group mean. This large variability is somewhat coherent with the 5 to 1 difference across labs testing English RSVP reading (Rubin and Turano, 1992; Latham and Whitaker, 1996; Fine et al., 1997, 1999; Chung et al., 1998; Pelli et al., 2007).

If high individual variability is the norm, the mean performance of any given sample would depend upon the actual proportion of fast and slow individuals. This is particularly the case for RSVP studies which are typically concerned with perceptual parameters and use a large number of trials but a small (often very small) sample size. In these conditions, variability between samples is expected to be quite high and this may substantially contribute to the very different reading rates reported in the literature.

Variability may be the diagnostic marker of the reading differences across regular and irregular orthographies. However, this first preliminary experiment had several pitfalls preventing any definite conclusion on whether the high inter-individual variability among English observers is a "real" phenomenon. Obviously, one possible source of variability would be the presence of a proportion of individuals with a reading deficit. All participants were self-reported proficient readers, but, given the absence of an independent evaluation using standardized reading measures, it is impossible to exclude such an explanation with certainty. Moreover, we did not have a measure of wpm in the case of words equated on number of phonemes (rather than letters). Based on these considerations it seemed important to confirm and extend the findings of Experiment 1 with a new group of subjects; this was carried out in Experiment 2. Additionally, it is unclear whether the difference in variability between the two groups is specifically related to the cognitive components involved in the performance with the RSVP or may be a more general phenomenon extending across reading tasks. This was the aim of Experiment 3.

EXPERIMENT 2: FUNCTIONAL READING ABILITIES AND RSVP READING SPEED

In this experiment we aimed to replicate Experiment 1 measuring RSVP reading speed in an independent sample. In order to exclude differences between samples related to more general cognitive efficiency and/or the presence of a reading deficit, standardized tests appropriate for the participants' age and language were administered to ensure that all participants were normal fluent readers. Additionally, the performance of English and Italian readers was examined both using lists of words matched for number of letters and lists of words matched for number of phonemes.

METHODS

Participants

Italian readers were 32 university students recruited from the student population of the Sapienza University of Rome; the English participants were 26 students recruited from the student population of the University of Hull. As shown in **Table 1**, the groups were matched for age ($t < 1$; $p = 0.59$); the years of schooling

Table 1 | Socio-demographic information and reading and Raven's SPM performance for the Italian and English samples of Experiments 2 and 3.

	Italian participants	English participants	Difference
Gender	15M, 17F	11M, 15F	$X^2 < 1$, p = 0.73
Mean age	23.8 (1.9)	23.0 (1.1)	T < 1, p = 0.59
Years of schooling	17.1 (1.6)	14.6 (0.5)	$t_{(56)} = 9.11$, p < 0.0001
Raven's SPM (mean standard score)	110 (9.6) (range: 93–128)	108 (10.4) (range: 90–140)	T < 1, $p = 0.82$
Word reading: errors (mean z score)	−0.24 (0.67)		
Word reading: speed -syllables/second- (mean z score)	−0.49 (1.03)		
Pseudo-word: reading errors (mean z score)	−0.38 (0.66)		
Pseudo-word: reading speed -syllables/second- (mean z score)	−0.11 (0.80)		
Word reading: TOWRE sight word efficiency (mean z score)		0.18 (0.54)	
Pseudo-word reading: TOWRE Phonemic Decoding Efficiency (mean z score)		0.65 (0.65)	

Unless otherwise specified, values in brackets indicate standard deviations.

were higher [$t_{(56)} = 9.11$, $p < 0.0001$] for the Italian than the English sample. These differences are presumably related to the longer Italian schooling system.

The following inclusion criteria were used to select the participants included in the two samples (English and Italian): (i) absence of neuro-sensory deficits or cognitive impairment (as assessed by Raven's Standard Progressive Matrices—SPM, Raven, 2008). (ii) Absence of a reading deficit assessed by single word and pseudo-word reading tests (for Italian: Martino et al., 2011; for English: the Test of Word Reading Efficiency—TOWRE, Torgesen et al., 1999); (iii) Normal or corrected to normal visual acuity; (iv) absence of a history of reading disorder. **Table 1** reports the performance obtained by the two language groups in the standard reading tests and the Raven's SPM.

Stimuli, apparatus, and procedure

Two lists of 80 stimuli were generated for each language, one consisting of words of 5 letters and one consisting of words of 5 phonemes. Again, words were all nouns: morphologically complex, archaic, obsolete, poetic and scientific words as well words with an opaque grapheme-to-phoneme correspondence or irregular stress were avoided in both languages. Words were selected from the MRC Psycholinguistic Database 2.0 (Wilson, 1988) for English and from the LEXVAR (Barca et al., 2002) and Colfis (Bertinetto et al., 2005) databases for Italian language. Note that for words selected by the Colfis database, values of n-size, imageability, age of acquisition and bigram frequency are computed with the same procedure used for the LEXVAR database (Barca et al., 2002). **Table 2** reports the values of frequency, number of letters and phonemes for each list. Note that for the English readers, the 5-letter list was the same as that used in Experiment 1.

The Italian and English lists were matched for frequency, n-size, imageability and age of acquisition (all $ps > 0.1$) but not bigram frequency (with a higher value of bigram frequency in the Italian relative to the English lists, according to MCWord database in English, and LEXVAR database in Italian; for the letter-matched list: $t_{(159)} = 4.63$, $p < 0.0001$; phoneme-matched list: $t_{(159)} = 5.20$, $p < 0.0001$. Also, lists were not matched for number of syllables since Italian words generally have higher values than English words [for the letter-matched list: $t_{(159)} = 14.41$, $p < 0.0001$; phoneme-matched list: $t_{(159)} = 4.96$, $p < 0.0001$]. Appendix B reports the means (and standard deviations) of the psycholinguistic variables for each experimental list.

Table 2 | Characteristics of Italian and English 5-letter and 5-phoneme words in Experiment 2.

		Italian	English
5-Letter words	No of phonemes	4.15 (0.36)	4.01 (0.70)
	Word frequency (mean)	29 (64)	25 (36)
5-Phoneme words	No of letters	5.45 (0.69)	6.03 (0.90)
	Word frequency (mean)	27 (88)	22 (24)

Note that the word frequency values were computed for Italian according to the Colfis database (Bertinetto et al., 2005; based on one million occurrences) and for English according to the CELEX database (Baayen et al., 1993). Values in brackets indicate standard deviations.

The apparatus and procedure were the same as in Experiment 1.

RESULTS

Differences between the two language groups on the log transformed reading rates were assessed through an ANOVA with language (English, Italian) as the unrepeated factor and list (letter-matched, phoneme-matched) as the repeated factor. The results indicated a significant main effect of the language factor [$F_{(1, 56)} = 13.46$; $p < 0.001$] with the English readers faster than the Italian ones on both the letter-matched list (geomean = 453 and *CI*: 344–598 for the English readers; 325 wpm and CI: 292–362 for the Italian readers) and the phoneme-matched list (geomean = 514 and *CI*: 407–649 for the English readers;

299 wpm and *CI*: 266–337 for the Italian readers). The main effect of list [$F_{(1,56)} < 1$, $p = 0.68$] and the language by list interaction [$F_{(1,\ 56)} = 2.7$; $p = 0.12$] were not significant. Results were replicated also when socio-demographic variables (i.e., gender, age, scholarity and Raven's SPM accuracy) were added to the analysis as covariates: only the main effect of the language factor was significant [$F_{(1,\ 55)} = 9.78$; $p < 0.01$]; no other main effect or interaction were significant, as well as the effect of the covariate variables.

Figure 2 shows the reading rates obtained by each individual for the two lists. As in Experiment 1, an inspection of the figure reveals much greater variability in the English than in the Italian sample. This is confirmed by the Levene's test for equality of variances (for the 5-letter list: $F = 12.20$, $p < 0.001$; for the 5-phoneme list: $F = 4.40$, $p < 0.05$). The English group contains both the fastest individual (reading at 1075 wpm) and the slowest individual (reading at 62 wpm).

Again, to control for the possible effect of introducing a nonlinear transformation on the difference in the variance of the two distributions, we made the same analysis in terms of untransformed threshold values. For the 5-letter list, mean duration thresholds were 112 ms ($SD = 38$) for the Italian group and 175 ms ($SD = 213$) for the English group. For the 5-phoneme list, duration thresholds were 123 ms ($SD = 48$) for the Italian group and 137 ms ($SD = 135$) for the English group. The Levene's test for equality of variances indicated that the variances of the two groups were different for the 5-letter list ($F = 15.88$, $p < 0.0001$) as well as the 5-phoneme list ($F = 6.94$, $p < 0.01$). These results indicate that the differences in variability between the two languages are genuine, i.e., they are not due to the use of a nonlinear transformation of data.

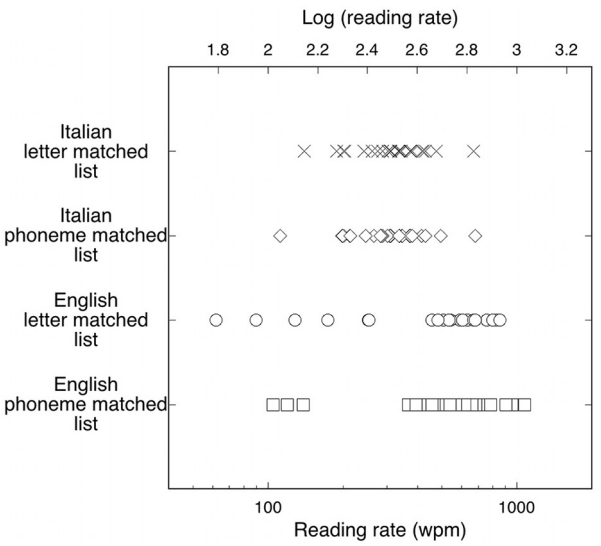


FIGURE 2 | Individual reading rates for Italian readers (Xs and diamonds for letter- and phoneme-matched lists, respectively) and English subjects (open circles and open squares for letter- and phoneme-matched lists, respectively). The upper scale shows the corresponding values expressed as a log (wpm). Reading rates for English observers are much more variable than rates for Italian observers.

Comments

Unlike Experiment 1, but in agreement with our informal observations based on children's data, the results revealed that English readers were on average faster than Italian readers. However, the results also replicated the much greater variability in the English sample compared to the Italian one, already observed in Experiment 1. In this experiment, we directly evaluated the subjects' reading proficiency in standard reading tests. Therefore, the large asymmetry in reading rates was present even after controlling for reading proficiency, indicating that differences in variability across languages may be a true phenomenon that needs to be explained.

Differences in average rate, absent in Experiment 1 and present in Experiment 2, may be interpreted, at least in part, as due to the sampling effect from a greatly variable distribution. Thus, as English individuals are likely to show extreme performance on both ends of the distribution, the relative proportion of such "fast" and "slow" individuals may greatly influence the general outcome of a study, unless a very large sample is examined. Note that in psychophysical studies using RSVP of stimuli, sample sizes are usually quite small and several experiments are actually run on very experienced observers (including experimenters).

One note of caution should be advanced in relation to the proficiency measures used in the two languages. We relied on standardized, validated procedures widely used in the two linguistic contexts. Clearly, it was not possible to use fully comparable instruments as different tasks and measures are traditionally used in the two clinical settings. Notably, English observers were all considered normal at a standard reading test which included an evaluation of speed; the TOWRE (Torgesen et al., 1999) uses a combined measure of speed and accuracy, based on the number of words (or pseudo-words) accurately read within 45 s. In this respect, it should be kept in mind that, in clinical testing, performance is measured with reference to typical samples, usually in terms of the standardized distance from the mean. So, large variability in the data would allow for greater distances from the mean to be accepted as normal. In other terms, there is no absolute way to establish normality other than in comparison to a group of individuals without apparent reading difficulties. So, if variability in reading speed is the norm, it will prove relatively difficult to be considered "atypical" in this particular measure. Finally, it should be kept in mind that, by limiting the influence of articulatory and eye movement components, the RSVP procedure allows for a much larger spread of measure than standard reading (which finds its upper limit with "rauding"; Carver, 1992).

In keeping with this last observation, we wondered whether the large variability in rates shown by readers of languages with irregular orthographies both here and in the literature is related to the characteristics of the RSVP paradigm or extends to other reading measurements. One widely used measure of reading speed is vocal RTs. They are usually measured to address the effect of lexical variables and build models that aim to explain reading. For example, the recent Connectionist Dual Process model (CDP++ model, Perry et al., 2010) simulates the effects found in reading aloud mono- and di-syllabic words and pseudowords, in stress assignment, regularity and syllable number. Furthermore, there is a large literature concerning the interpretation of RT

measurements and, in particular, there are models developed to account for individual differences in this measure. According to Wagenmakers and Brown (2007), the three general characteristics of RT distributions that need to be accommodated by any model are that: (a) RT distributions are typically skewed to the right; (b) this skew increases with test difficulty; and (c) the spread of the distribution grows as a function of the mean. Indeed, various models have been developed to tackle this last question, i.e., to account for the relationship between the mean and the standard deviation of a response time distribution (e.g., Faust et al., 1999; Myerson et al., 2003), and they can be particularly fruitful in the present context. Thus, with the aim of addressing the selectivity of the variability effect across languages, in Experiment 3 we extended our observations to vRT measurements.

EXPERIMENT 3: EXAMINING THE MEAN AND THE STANDARD DEVIATION

Despite the emphasis given by most models of reading on the prototypical reader, there is clear evidence that variation in reading skills uncover the underlying process of reading (Balota and Spieler, 1999). To date, the systematic study of individual differences in RT measures has been particularly focussed on aging (Salthouse, 1985; Cerella, 1990) and practice (e.g., Logan, 1992) effects and much less so on understanding the performance of young proficient adults.

One line of investigation has focussed on the possible modulating role of general speed of processing differences across groups. It has been noted that, to fully investigate selective effects of experimental manipulations (e.g., frequency effect), the global factor influencing the differential speed of processing across groups must be taken into account (Faust et al., 1999). Studies have found that more difficult conditions (e.g., low frequency long words compared to high frequency short words) produce larger differences in generally slower groups of individuals (e.g., older adults) due to over-additive interactions (Salthouse, 1985; Cerella, 1990; Myerson et al., 1992; Faust et al., 1999). In line with the presence of an over-additivity effect, Paulesu et al. (2000) reported a larger lexicality effect (words read faster than pseudo-words) in the generally slower sample of English readers than in their sample of Italian readers (who were faster readers). In a developmental perspective, Zoccolotti et al. (2009) found a lexicality effect from grade 1 to grade 8; however, the effect increased progressively with age, when the role of over- additivity was controlled for. In this vein, we will compare vRTs as a function of task difficulty (manipulating a variable such as word frequency) between two languages that, as we have seen in Experiments 1 and 2, differ in terms of mean performance (as well as in variability). One aim of the experiment was to assess the relationship between individual differences between groups and the role played by a lexical variable (i.e., word frequency).

A second line of investigation focussed on the characteristics of the distribution of vRTs. There is a large literature that examined which is the most appropriate distribution to describe the typical skew observed with RTs. In this vein, possible candidates are the ex-Gaussian, the shifted lognormal, the shifted Wald, the shifted Weibull, and the Gumbel distribution (for a discussion among these options see Wagenmakers and Brown, 2007). Yap et al. (2012) extensively investigated individual differences in the reading performance of young English adults in relation to vocabulary knowledge by applying the ex-Gaussian analysis (i.e., a convolution of a Gaussian and exponential distribution) to investigate the RT distributions in reading (Ratcliff, 1979). Interestingly, individual differences were associated with diverse distributional patterns and cognitive abilities (Yap et al., 2012). In particular, results emphasized the role of stable lexical processing characteristics at the individual level. Interestingly, different ex-Gaussian parameters were differentially sensitive to lexical knowledge; thus, the correlation between vocabulary knowledge and vRTs was greatest for the parameter (τ) expressing the exponential component (particularly sensitive to the tail of the RT distribution). Following these observations, we will apply the ex-Gaussian analysis to investigate the RT distributions and the modulating role of a lexical variable such as word frequency in reading of English and Italian proficient readers.

Finally, an interesting line of research on RTs is the development of general models that try to understand the individual performance by decomposing this measure into its constituents. For example, in explaining the relationship between task difficulty (expressed as average speed) and individual differences (measured by SDs), Myerson et al. (2003) proposed the DEM, a two-compartment model. Accordingly, the observers' response is related to a sensory-motor compartment that is generally invariant across subjects (including fast and slow populations, such as old and young adults), and a cognitive compartment that determines how individual differences vary as a function of task difficulty. Critically, the DEM envisages specific predictions to evaluate the relative contributions of the two compartments. These predictions will be tested in the present sample of English and Italian readers.

The general aim of the experiment was to extend the RSVP results to the vRT measures and to assess individual differences within and between groups and the role played by a lexical variable (i.e., word frequency). In examining these questions we took advantage of the previous general literature on RT measures. Thus, we examined (a) the possible presence of over-additivity effects; (b) the distribution of vRTs by ex-Gaussian analysis; and (c) the fit of vRT measures to the DEM (Myerson et al., 2003).

METHODS

Participants

The same participants in Experiment 2 also took part in Experiment 3.

Stimuli, apparatus, and procedure

Stimuli used for the vRTs experiment were selected from the two lists of 80 words used in each language for the RSVP experiments. A sub-list of 20 high-frequency words and one of 20 low-frequency words was created from both the 5-letter and 5-phoneme lists (see **Table 3** for a description of the lists).

In Italian and English (for both the 5-letter and 5-phoneme stimuli), the sub-lists of high- and low-frequency

Table 3 | Characteristics of Italian and English 5-letter and 5-phoneme (high- and low-frequency) words in Experiment 3.

			Italian	English
5-Letter words	Low frequency words	No of phonemes	4.25 (0.44)	4.00 (0.76)
		Word frequency (mean)	3 (1)	3 (2)
	High frequency words	No of phonemes	4.05 (0.22)	4.27 (0.80)
		Word frequency (mean)	57 (57)	61 (19)
5-Phoneme words	Low frequency words	No of letters	5.70 (0.69)	6.33 (1.05)
		Word frequency (mean)	3 (4)	3 (1)
	High frequency words	No of letters	6.00 (0.82)	6.00 (0.85)
		Word frequency (mean)	55 (79)	64 (15)

Note that words frequency values were computed for Italian according to the Colfis database (Bertinetto et al., 2005; based on one million occurrences) and for English according to the CELEX database (Baayen et al., 1993). Values in brackets indicate standard deviations.

words did not differ for imageability, number of letters, phonemes, N-size and bigram frequency (all $ps > 0.1$), but differed for age of acquisition (as expected due to the high correlation with frequency). The four sub-lists (5-phoneme high-frequency words, 5-phoneme low-frequency words, 5-letter high-frequency words, and 5-letter low-frequency words) in English did not differ from the Italian sub-lists for any variable considered, except for the number of syllables, which were higher in Italian than in English for the 5-letter sub-lists. The means (and standard deviations) of these psycholinguistic variables are reported for each set of experimental stimuli in Appendix C.

Participants were seated ca. 57 cm from the computer screen. Stimuli were presented using the E-prime 2 software. Each trial began with a fixation point that remained on the screen for 500 ms. Subsequently, a word appeared in the same position. The stimulus remained on the screen until the participant responded. High and low frequency words were randomized for each participant and presented in separate blocks. The order of presentation of the two blocks was balanced across subjects. Five practice stimuli preceded each block. The participant was requested to read the stimulus as quickly and accurately as possible. VRTs were recorded using a voice key (S-R Box). The computer recorded the onset of the vocal response. The experimenter manually recorded pronunciation errors. The responses were tape-recorded to allow offline rechecking. The vRTs corresponding to errors were excluded from the analyses. Self-corrections and wavers were considered errors and the corresponding vRTs were not included in the analyses. Invalid responses (due to technical problems) and vRTs below 200 ms were also excluded from the analyses (1.8% in the English sample and 2.0% in the Italian sample).

RESULTS

Frequency effect as a function of language

Table 4 reports the means (and standard deviations) of vRTs and error rates of Italian and English participants in each condition of the experiment. As it can be seen from the table, the percentage of errors was very low for both groups and, so, no formal analysis of error measures was made. The results on vRTs were submitted to an ANOVA with language as the unrepeated factor, and list (letter- and phoneme-match) and frequency (high and low) as repeated measures.

The main effect of frequency was significant [$F_{(1,\ 56)} = 86.59$; $p < 0.0001$]: low-frequency words were read slower (511 ms) than high-frequency words (482 ms). No other main effects or interactions were found to be significant: vRTs of the two groups did not differ (English observers: 491 ms, Italian observers: 502 ms; $F < 1$, $p = 0.50$), and the two lists were equivalent in terms of reading speed (letter-match: 496 ms, phoneme-match: 497 ms; $F < 1$, $p = 0.95$). Thus, in the absence of a general speed difference between the two linguistic groups, the effect of word frequency was present but did not vary between the two language groups. Results were replicated also with socio-demographic variables (i.e., gender, age, years of schooling and Raven's SPM accuracy) added to the analysis as covariates: only the main effect of frequency was significant [$F_{(1,\ 55)} = 9.32$; $p < 0.01$]; no other main effect or interaction was significant, as well as no effect of the covariate variables.

Individual difference distribution as a function of language

We characterized the vRT distributions of English and Italian participants in terms of the ex-Gaussian probability density functions. The ex-Gaussian distribution is the convolution of a Gaussian (normal) and exponential distribution that accounts for the positively skewed RT distribution often seen in empirical data. We used the MatLab analysis tools provided by Lacouture and Cousineau (2008) and applied the following ex-Gaussian function:

$$f(x|\mu, \sigma, \tau) = \frac{1}{\tau} \exp\left(\frac{\mu}{\tau} + \frac{\sigma^2}{2\tau^2} - \frac{x}{\tau}\right) \Phi\left(\frac{\mathrm{x} - \mu - \frac{\sigma^2}{\tau}}{\sigma}\right) \quad (1)$$

in which the exponential component (exp) is multiplied by the cumulative Gaussian component (Φ). The resulting ex-Gaussian distribution contains three parameters: mu (μ) and sigma (σ) are the mean and standard deviation of the Gaussian distribution, and tau (τ) is the mean of the exponential component. We estimated the three parameters values of the individual participants' data distributions across items by applying the maximum likelihood procedures described by Lacouture and Cousineau (2008). Appendix D presents the individual mean vRTs, its standard deviation, as well as the ex-Gaussian parameters for individual participants across all conditions (letter- and phoneme-matched, high- and low-frequency lists).

Table 4 | Means (and standard deviations) of vRTs and error rates (% of errors) of Italian and English participants for all experimental conditions of Experiment 3.

Words		vRTs (ms)				Errors (%)			
		Italian participants		English participants		Italian participants		English participants	
		Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
5-Phonemes	High frequency	494.7	51.5	474.4	62.1	0.9	0.3	0.8	0.4
5-Phonemes	Low frequency	514.4	65.6	507.4	82.3	0.5	0.3	0.4	0.3
5-Letters	High frequency	488.0	61.3	476.4	63.9	1.3	0.4	0.6	0.4
5-Letters	Low frequency	522.7	71.1	505.8	79.0	0.6	0.3	0.2	0.3

Table 5 | Means (and standard deviations) for the ex-Gaussian parameters of Italian and English participants across experimental conditions of Experiment 3.

Ex-Gaussian parameters	Italian participants		English participants		Student test		Levene test	
	Mean	*SD*	Mean	*SD*	*t*	*p*	*F*	*p*
Mu	439.0	45.9	445.6	59.7	−0.46	0.64	0.36	0.55
Sigma	35.7	18.7	75.9	32.3	−5.63	< 0.0001	8.67	0.005
Tau	66.7	28.0	45.3	40.3	2.30	0.026	7.42	0.009

Group comparisons (by Student t-test) and Levene's test for equality of variances are presented.

In keeping with the analyses carried out in Experiments 1 and 2, we examined the various ex-Gaussian parameters both in terms of group differences (by means of *t*-tests) and in terms of equality of variances (by means of Levene's test). Means (and SDs) and results of these analyses are presented in **Table 5**.

T-test comparisons revealed a significant difference between the standard deviation (σ) of the Gaussian component but no difference in the mean: Italian observers showed a μ of 439 ms and a σ of 36, while English observers a μ of 446 ms and a σ of 76. Furthermore, the τ (representing the mean of the exponential component) was significantly larger in the Italian (67 ms) than in the English (45 ms) group.

The Levene's test for equality of variances indicated that the variances of the two groups were not different in the case of the μ parameter while they were significantly different for the σ and τ parameters, in both cases indicating greater individual variability in the English than in the Italian sample (see **Table 5**).

To summarize these results: (a) the two linguistic groups were similar in μ both in terms of mean performance and inter-individual variability; (b) Italian observers showed higher τ and lower σ values than English observers; and (c) independent of group mean differences, English observers were more variable across individuals for both τ and σ, but not μ.

Figure 3 presents the fits of the ex-Gaussian functions across participants to the empirical data separately for the two linguistic groups, using a super-subject approach (as in Balota and Spieler, 1999). In agreement with the individual data, the results of the fit of the Italian data resulted in a μ of 416 ms, a σ of 49, and a τ of 89, while the English fit indicated a μ of 412 ms, a σ of 89, and a τ of 78. As shown by the figure, the larger τ obtained by

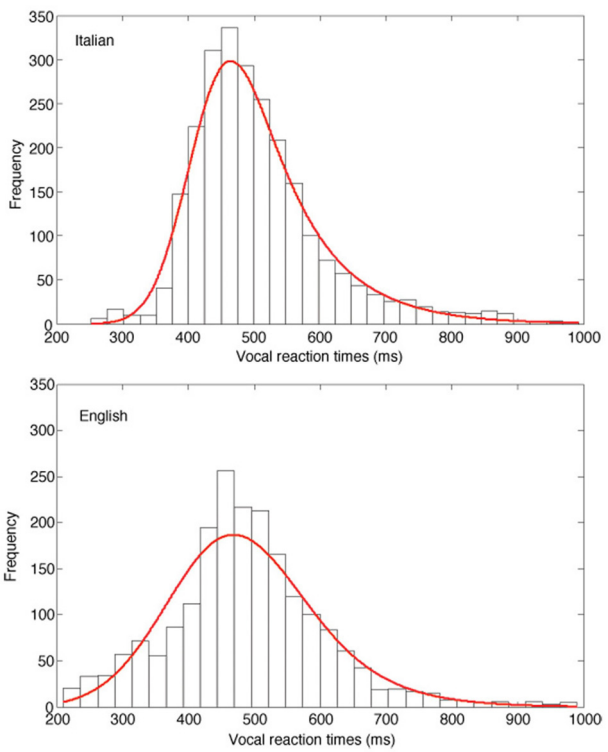


FIGURE 3 | Frequency distribution of vRTs of the Italian (upper panel) and the English (lower panel) sample. The red solid lines represent the best fit of the ex-Gaussian distributions of the data obtained by the two samples.

Italian readers is evident in the positively skewed vRT distribution (upper panel) relative to the English data distribution (lower panel), while a larger variability characterizes this latter group.

In order to clarify the relation between the distributional parameters and the lexical status of the stimuli we applied the ex-Gaussian fit separately for the two lists of high and low frequency words. It must be noted that due to the limited number of items (40 for each participant) parameter estimates may only be taken as a suggestion of an existing relationship. Three separate ANOVAs were applied to the resulting parameters with group (English and Italian) as unrepeated factor and frequency (high and low) as repeated factor. The results for the τ estimates revealed the significant main effect of group and frequency [respectively: $F_{(1,\ 56)} = 8.04$, $p < 0.01$; $F_{(1,\ 56)} = 22.55$, $p < 0.0001$] but not of their interaction ($F < 1$, $p = 0.79$). The relative means for the τ parameter were 40 for the English group and 65 for the Italian, and 41 for high- and 67 for low-frequency words. As for the σ parameter only the main effect of group was significant [$F_{(1,\ 56)} = 37.26$, $p < 0.0001$], with a larger value for the English sample (72.5) relative to the Italian (34.5). The main effect of frequency and its interaction with group were not significant ($F < 1$, $p = 0.87$ and $F < 1$, $p = 0.77$, respectively). No main effect (group: $F < 1$, $p = 0.53$; frequency: $F < 1$, $p = 0.63$) or interaction was significant in the μ component of the distributions ($F < 1, p = 0.97$).

The results indicate that the τ (but not the μ and σ) parameter is sensitive to the effect of frequency. As in the general analyses presented above, larger τ values were present for Italian individuals but no differential effect of language on the frequency effect was detected.

Modeling vRTs as a function of language

In order to investigate the nature of the individual differences between the two groups we applied the DEM (Myerson et al., 2003). According to this model, the slope of the linear relationship between mean vRTs and SDs is indicative of the amount of processing required by the observers to perform the task and it directly assesses the cognitive compartment (i.e., the slope indicates the correlation between the cognitive stages involved in the task). By contrast, the intercept on the x-axis of this linear relationship estimates the time of the non-decisional sensory-motor compartment (which is supposed to be invariant across observers). Thus, the DEM allows for independent estimations of the cognitive and sensory-motor components in determining individual differences in task performance.

Figure 4 shows the relationship between individual SDs and vRTs in the two groups: data for Italian observers are presented on the left, while those of English observers are shown on the right. The variability grows linearly with increasing condition means for Italian readers; this is less clear for English readers. The solid line in **Figure 4** represents the DEM prediction calculated on all the participants using the following equation (Equation 2):

$$SD = \left(r - \frac{\sigma_c}{\alpha}\right)(RT - t_e) \quad (2)$$

where σ, α, t_e, and r are parameters that are free to vary and represent the variance and amplitude of the effects, the time required by the sensory-motor compartment, and the theoretical correlation between the cognitive stages, respectively. In **Figure 4**, the sensory-motor compartment is represented by the x-intercept of the regression line. In the case of the Italian sample, the model explains relatively well the variability in the data ($R^2 = 0.42$) with an estimated time for the sensory-motor compartment of 239 ms and a slope of 0.30. Note that these values are close to those typically reported in the literature (Myerson et al., 2003). By contrast, the model does not account well for the English data ($R^2 = 0.10$). The slope relating means and SDs is nearly flat (0.17) and no reliable estimation of the sensory-motor compartment is possible; indeed, the x-intercept of the regression line is negative (−26 ms).

RSVP and vRT reading comparison

As the same subjects participated in Experiments 2 and 3 this allowed examining the consistency of individual differences between the RSVP (in terms of the log of wpm), and the vRT measures. The Pearson correlation between the two measures was 0.35 for the Italian participants ($p < 0.01$) and 0.31 for English participants ($p < 0.05$). The reading measures in Experiments 2 and 3 differ in terms of absolute performance level and of the response compartment involved. RSVP reading thresholds are calculated at a criterion level of task performance of 80%, while vRTS are measured for correct responses (ideally 100% correct). In addition, while RSVP maximizes the decoding component of the process leaving unlimited time to utter the word, vRT measures include the programming and the beginning of the motor execution (for the role of motor compartment on vRT and total time measures see Martelli et al., 2014). Nonetheless, the analysis shows that the two measures are significantly correlated.

Comments

Do the reading skills measured by the speed in vRTs interact with the size of the lexical effects? Estimation of reading skills did not reveal a mean group difference between a regular and an irregular orthography in adult proficient readers. In this context, standard analyses based on mean performances did not show a significant interaction between frequency and language, indicating a similar use of the stimulus lexical properties by the two groups.

The analysis of the ex-Gaussian probability density functions revealed that language differences are captured by the differential weight of the two components (Gaussian and exponential) in determining the vRT distributions of the two linguistic groups: Italian observers showed higher τ and lower σ values than English observers while no group difference was detected in the case of μ values. As a consequence, examining differences across regular and irregular orthographies only with reference to the mean fails in capturing the phenomenon. Additionally, the results indicate that τ, but not σ and μ, are modulated by the lexical status of the stimuli. These findings are in keeping with previous data from Yap et al. (2012) who reported that vocabulary knowledge was correlated to τ more highly than to μ in speeded pronunciation of words (as well as in a lexical decision task). More generally, similar results have also been reported for decision and selective attention tasks (Schmiedek et al., 2007; Tse et al., 2010).

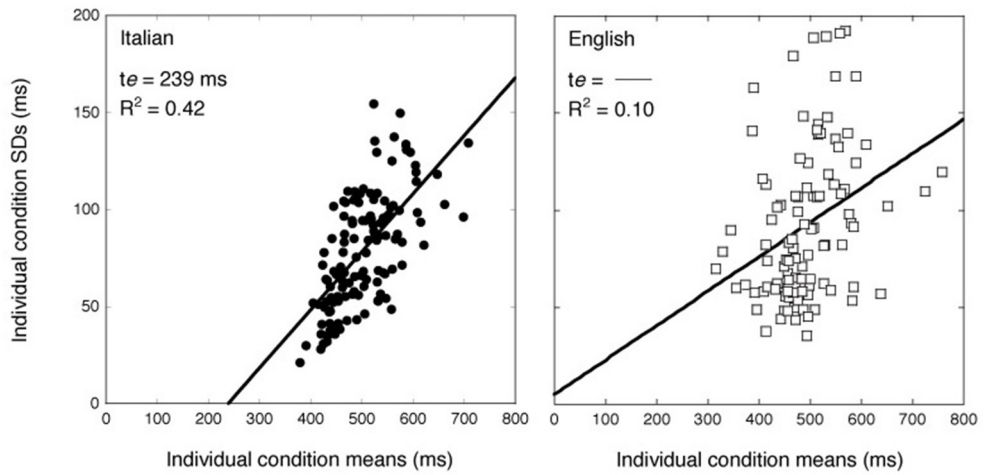


FIGURE 4 | Individual participants' SDs over items as a function of mean vRTs. On the left Italian readers; on the right, English participants. The solid line represents the best fit of the DEM to the data (Myerson et al., 2003).

A critical interest of the present study is to examine the possible presence of inter-individual differences between the two linguistic groups. Results indicated a profile of individual differences that varied as a function of the ex-Gaussian parameters: English observers were more variable for τ and σ but not for μ. In comparing data from this experiment to the findings of Experiments 1 and 2 note that in the case of RSVP paradigm only mean performance values are available. So, in this vein, one should note the differential outcome with the two paradigms: in the case of the RSVP, one obtains a different spread of performances with the English sample containing the fastest and slowest individuals. In the case of vRTs, this differential spread is not present (as indicated by the pattern of data in the case of μ values). However, English observers were more variable both in terms of σ and τ. The former finding indicates greater intra-individual variability (further comments will be advanced in the general discussion). Of particular interest is the greater variability in τ as this parameter is the one that selectively captures the effect of the lexical status of the stimuli.

One of the most basic results in the RT literature is that slower vRTs are accompanied by higher variability (Wagenmakers and Brown, 2007). This is commonly true of both group and condition comparisons. So, older people are generally slower and more variable than younger individuals; more difficult conditions are invariably associated with larger variability values (Faust et al., 1999; Myerson et al., 2003). By contrast, in the current study, English participants were more variable but not slower than Italian participants across conditions. So, the pattern shown by English readers is at odds with the basic predictions of models interpreting global effects of performance. Note that here we are showing individual observer's means and SDs separately for each condition. Myerson et al. (2003) DEM is typically applied to condition means segregating slow vs. fast subjects (e.g., comparing the fast to the slow quartile of the observers distributions) in studies comprised of several independent measures. Application of DEM indicates that fast and slow subjects are described by the same relationship between means and SDs. The consequence of these linear relations is that the difference between the vRTs of the subgroups of fast and slow processors increases proportionally with the average vRT as SDs do, so that the data for the two groups are typically fit by the same regression line. Thus, under the assumption that for all observers the same processing steps are recruited by the task and that speed of processing affects all the steps equally (proportionally) the model predicts the same relationship (same slope) between SDs and means at an individual and at a group level.

Our results indicate that the DEM does not adequately fit the data of English readers (Myerson et al., 2003). Notably, this model has been largely developed on experiments run on English samples, although typically not on reading tasks (Myerson et al., 2003). So, the differential outcome may indeed be specific to reading.

DISCUSSION

Are group means effective estimates of reading speed? Results of Experiments 1 and 2 strongly indicate that the reliable difference between the two language groups is expressed by the variability in the distribution of performances rather than by their mean. Studies with the RSVP reading speed in English participants have been unable to clearly ascertain a value for reading speed, producing speeds that differ in wpm up to a factor of 5 (Rubin and Turano, 1992; Latham and Whitaker, 1996; Fine et al., 1997, 1999; Chung et al., 1998; Pelli et al., 2007). Differences across studies may be partially explained by the diverse procedures adopted: presence, or absence, of a mask preceding and following the stream of words in the trial, presence, or absence, of context (i.e., random words vs. sentences), number of words presented in a trial (for an overview on the effects of these variables in RSVP reading see Primativo et al., in preparation). However, our results indicate that part of the differences in reading speed estimates obtained by different laboratories may be related to sampling biases. English readers are much more

variable; thus, sample size and selection criteria greatly affect the reliability of the mean in defining the group speed (especially for the small sample sizes typical of these studies). Our results indicate that differences in speed across labs may be in part reconciled in light of the large variability shown by the English population.

Also results with vRTs (Experiment 3) point to the presence of greater differences in variability between English and Italian observers than in terms of mean performances. However, results in this case indicate that different parameters capture the group differences in variability: English observers were more variable for σ and τ values but not for μ, i.e., the mean of the Gaussian component.

How can these differences be accommodated? One possible interpretation is that, by minimizing the role of memory, pronunciation time and eye movements, performance in the RSVP paradigm closely captures the efficiency in decoding; so, individual differences are directly reflected in differential ranges, with the English sample containing both the slowest and fastest individuals. In the case of RT measures, the available literature indicates a more complex relationship between performance and decoding. There is a consensus that RT measures contain both decisional and non-decisional components although there are different approaches to separate them (e.g., diffusion model: Ratcliff et al., 2004; DEM: Myerson et al., 2003). Within the DEM to which we refer here, RTs are a compound of a sensory-motor compartment and of a decisional compartment. Myerson et al. (2003) propose that individual differences are confined to the decisional component of the response. In this perspective, it is not surprising that individual differences are not well captured by variations in the mean as this expresses both decisional and non-decisional components of the response. To provide a general frame for this distinction consider that, based on DEM, in the present experimental conditions the sensory-motor and cognitive compartments each account for about half of the processing time in the Italian sample. Thus, 239 ms was the estimate for the former compartment; subtracting this value from the overall mean (502 ms), we obtain an estimate of 263 ms for the cognitive compartment. Note that the two compartments were not distinguishable among English observers (see further comments below).

Furthermore, in keeping with the idea that the typical skew to the right of RTs increases with test difficulty (Wagenmakers and Brown, 2007), we observed that τ values captured changes in performance as a function of the lexical status of the stimulus better than μ values. This pattern is consistent with previous observations on both reading (Yap et al., 2012) and non reading (Schmiedek et al., 2007; Tse et al., 2010) tasks. Accordingly, we found that English observers were more variable in their τ values. So, within this reasoning, the outcome of the three experiments can be reconciled by stating that English observers showed greater individual differences than Italian observers in the parameter which, in each paradigm, is sensitive to variations in task difficulty (lexical status in our case), i.e., mean values in the case of the RSVP and τ values in the case of the vRTs. Note that in the case of vRTs, English observers were more variable than Italian observers also in terms of the variability of the gaussian component of the response (σ). Further comments will be made on this point when commenting the results within the DEM model.

Is there a processing speed difference in reading in regular and irregular orthographies once (most) cognitive variables are taken into account to match stimuli across languages? Experiment 2, but not Experiments 1 and 3, showed that English readers were faster than Italians. However, in all three experiments the English observers were more variable (although on different critical parameters). English and Italian differ in the degree of consistency in the mappings of letters onto sounds as well as in the complexity of the syllabic structure. The lower syllabic complexity of Italian language enables for easy segmentation of words into phonemes/syllables and, in turn, to effectively acquire grapheme-to-phoneme mappings. On the other hand, in English, the embedding of grapheme-phoneme correspondences in consonant clusters makes it more difficult to acquire these correspondences. In fact, Seymour et al. (2003) found that syllabic complexity affects accuracy and speed of reading nonwords (although not familiar words) and exaggerates the lexicality effect. Moreover, it has been suggested that the preferred grain size unit (i.e., the number of graphemes and phonemes) of the lexical entries differ across languages and determines different developmental constraints as well as the characteristics of adult fluent reading (Ziegler and Goswami, 2005). Ziegler et al. (2001) compared word and pseudo-word reading of German (a regular orthography) and English participants reading identical words (words written identically in the two languages such as ball, park, and hand) as a function of their length. Results showed that reading 5- and 6- letter words, the German participants were about 50 ms slower than the English sample. Conversely, Paulesu et al. (2000) found that adult Italian readers were faster at recognizing both words and pseudo-words relative to English readers. Frith et al. (1998) measured reading accuracy and speed of German and English children ranging in age between 7 and 9 years. Interestingly, they found that on average English children read at a slower speed and less accurately than German children, also showing a larger lexical effect. However, selecting a subgroup of "good readers" that made no errors in the easy items, they found English children to be slightly faster than their German peers. The results of the present experiments strongly indicate that English readers are more variable, and that the group mean *per se* fails to capture the phenomenon of the differences across languages. Thus, it is possible that individuals read a language with opaque orthography and complex syllabic structure adopting different processing strategies each contributing to reading with differential efficiency.

One of the aims of this study was to investigate the relationship between the general speed factor and the efficiency of the orthographic decoding on vRTs. Indeed, larger effects of the experimental manipulations are expected in the case of differences in overall processing time across individuals (i.e., larger effects for slower individuals, Faust et al., 1999). In the absence of a general speed difference in vRTs (Experiment 3) no over-additive group interactions are expected. Nonetheless, some data obtained by our research group on English and Italian children in reading single words and pseudo-words may be relevant on this issue (Marinelli et al., submitted). We found that, contrary

to the prediction of a larger RT variability in slower individuals (Faust et al., 1999; Myerson et al., 2003), the English sample was generally faster but more variable than the Italian sample across conditions, providing additional evidence that increased variability is a specific characteristic of English readers. Large inter-individual differences have also been found in other studies with English children, both when the English readers were faster (Ellis and Hooper, 2001; Ellis et al., 2004) or slower (Patel et al., 2004) than readers of regular orthographies. Taken together the results of the three experiments show that large variability is not associated with slower speed in the case of the English sample. This highlights the importance of examining the shape of the distributions to understand the underlying phenomenon.

Do the individual differences across languages arise from different strategies adopted in reading? One source of individual difference could arise from readers emphasizing different strategies or types of information during reading. Yap et al. (2012) linked the distributional characteristics with the dynamics of information accumulation over time. They found that a fluent lexical process, measured by good vocabulary knowledge, was associated with more efficient accumulation of information and lower τ. Accordingly, we found that only τ, and not σ or μ, were modulated by the lexical status of the stimuli (i.e., word frequency). If small τ is associated with higher use of the lexical strategy of reading (Yap et al., 2012), this may be more pronounced in the English population (i.e., lower τ indicates a more efficient process). However, word frequency modulated the τ parameter in a similar way in both the English and Italian samples. Again, the insensitivity in detecting mean group differences may be linked to the presence of large individual differences; so, apart from showing lower τ values, English observers were also significantly more variable in this parameter. Therefore, we feel that the possibility that the lexical strategy of reading is more pronounced in the English may require further testing before such hypothesis can be confidently rejected.

The DEM assigns the difference between individuals to the amount of cognitive processing required by the task predicting the relationship between mean and SD (Myerson et al., 2003). Applying the DEM to the data of Experiment 3 revealed that, in the case of the English sample, the SDs were not linearly related to the mean vRTs. There is a large body of literature that builds on the relationship between mean RTs and SDs to account for individual differences across slow and fast groups (e.g., Bashore and Ridderinkhof, 2002; Myerson et al., 2003). These studies investigate different cognitive processes, ranging from recognition to counting (Cerella et al., 1980; Cerella, 1985; Logan, 1992; Mayr and Kliegl, 1993; Hale and Jansen, 1994; Zheng et al., 2000; Palmer et al., 2011). The relationship between the standard deviation and the mean of the RT distribution highlights a general rule (Wagenmakers and Brown, 2007) that must be taken into account when looking for selective effects (Hale and Jansen, 1994; Faust et al., 1999; Zheng et al., 2000; Myerson et al., 2003). Indeed, most models of reading are based on the selective effects of lexical variables (e.g., CDP++ by Perry et al., 2010). Our results indicate that this relationship does not hold for reading speed in English. Note that, in a counting task, English participants show the expected relationship (Logan, 1992) but, as shown here, this is not the case for reading. Therefore, this makes a special case for English individuals and the reading task.

It is difficult to understand why the general linear law between means and standard deviations does not hold in this particular instance. Wagenmakers and Brown (2007) identify three boundary conditions under which no linear relationship between means and standard deviations is expected, i.e.: (a) manipulations affecting non-decision times; (b) mixtures (i.e., two different decisional processes going on at the same time) and (c) serial and exhaustive processing. An example of a mixture is when a task (e.g., counting dots) is solved in a moment of transition between the use of an algorithm (typically used in the early stages of learning) and of an automatic retrieval strategy (as in the case of the instance theory; Logan, 1992). Reading models can be seen in this perspective. Thus, at least to some extent, the dual route model (Coltheart et al., 2001) appears compatible with the instance theory, in that the sub-lexical route relies on an algorithm and the lexical route activates individual traces from the orthographic lexicon (i.e., specific instances). However, it seems extremely unlikely that English proficient adults are in a moment of transition between the two routes (if anything, this interpretation could apply more easily to Italian readers who supposedly develop their lexicon more slowly). The third boundary condition also does not seem to apply to the present data; it seems unlikely that reading is carried out through serial and exhaustive processing. At any rate, were this the case, one would expect means to vary linearly as a function of variances, rather than SDs (Wagenmakers and Brown, 2007). However, the results for English readers shown in **Figure 4** remained the same when we used variances instead of standard deviations, suggesting, as expected, that failure for linearity between means and SDs is not related to the task involving a serial and exhaustive processing.

Evaluating the first boundary condition (i.e., manipulations affecting non-decision times) identified by Wagenmakers and Brown (2007) is more complex. In general, models do not predict a relationship between means and standard deviations for the non-decisional component of the response. This is the case for the diffusion model (Ratcliff, 1979, 2002) as well as for the DEM (Myerson et al., 2003) to which we refer here. Variations among individuals and tasks certainly cannot be simply viewed as due to differences in sensory processing and motor preparation processes. However, in evaluating this boundary condition, it should also be considered that any systematic bias in the modulation of the response (such as response conservativeness) would be incompatible with the linear law (Wagenmakers et al., 2005). In this perspective one may consider the time criterion account for naming speed advanced by Lupker and colleagues in a number of studies carried out on English speaking individuals (e.g., Lupker et al., 1997; Taylor and Lupker, 2001; Kinoshita and Lupker, 2002; Chateau and Lupker, 2003). According to this hypothesis, participants set a point in time at which they try to respond to all stimuli in a given block. When easy and hard stimuli are mixed, the placement of the time criterion is intermediate compared with that in pure blocks of easy and hard stimuli; thus, responses to easy stimuli slow down and responses to hard stimuli speed up (thus altering the relationship between speed of

response and task difficulty which is at the base of the linear law). Notably, recent evidence (Paizi et al., 2010) indicates that a time criterion account does not easily explain reading RTs in Italian young adults. Thus, unlike what is typically reported in English-speaking individuals, word frequency effects are independent of list context manipulations (Paizi et al., 2010). So, one possibility is that the atypical pattern in the relationship between means and standard deviations is due to the fact that English readers (more so than readers of a regular orthography) refer to a time criterion when they try to read under speeded time conditions.

This hypothesis may also be instrumental in understanding the difference in intra-individual variability observed between the two linguistic groups: English observers present much higher σ values. Presumably, these values in part reflect the degree of noise in the data (whether arising from decisional or non-decisional components of the response). In this vein, it is interesting to note that σ values are extremely small in Italian observers, averaging 35.7 ms, with also very little inter-individual variation ($SD = 18.7$). If we assume that only English observers superimpose a time criterion on their response it becomes reasonable to imagine an increase in their individual intra-trial variability (independent of mean changes). Indeed, adopting a time criterion modifies the pattern of individual response in a way which, on the one hand, does not appreciably modify mean performance and, on the other, is inherently symmetrical and, as such, at least compatible with a Gaussian distribution. As the time criterion reflects a selective bias in the response, this perspective would also help understanding why σ values are insensitive to the lexical properties of the stimulus. Overall, we propose that the increase in intra-individual variability shown by English observers might be interpreted as due to a combination of two factors, intra-trial noise and reference to a time criterion for setting up the response. Only the former factor would be active in the case of individuals reading a very regular orthography, such as Italian. Clearly, further *ad hoc* research designs are needed to fully evaluate this interpretation.

Most universal models of reading and reading acquisition are based on mean vRTs of English participants as a function of lexical manipulations (e.g., Seidenberg and McClelland, 1989; Coltheart and Rastle, 1994; Coltheart et al., 2001; Perry et al., 2007, 2010). The current results question the appropriateness of building a universal reading model just on the very language group who in reading performance does not conform to the general predictions of models of RT performance.

ACKNOWLEDGMENTS

We would like to thank Caterina Di Serio, Maria Ferrara, Alessia Rossetti, and Roberta Zanonzini for their help in data collection. This work was supported by grants from the Italian Department of Health and Sapienza University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2014.00903/abstract

REFERENCES

Baayen, R. H., Piepenbrock, R., and Van Rijn, H. (1993). *The CELEX Lexical Database (CD-ROM)*. Philadelphia, PA: University of Pennsylvania, Linguistic Data Consortium.

Balota, D. A., and Spieler, D. H. (1999). Word frequency, repetition, and lexicality effects in word recognition tasks: beyond measures of central tendency. *J. Exp. Psychol. Gen.* 128, 32–55. doi: 10.1037/0096-3445.128.1.32

Barca, L., Burani, C., and Arduino, L. S. (2002). Word naming times and psycholinguistic norms for Italian nouns. *Behav. Res. Methods Instrum. Comput.* 34, 424–434. doi: 10.3758/BF03195471

Barlow, H. B. (1962). Measurements of the quantum efficiency of discrimination in human scotopic vision. *J. Physiol.* 160, 169–188.

Bashore, T. R., and Ridderinkhof, K. R. (2002). Older age, traumatic brain injury, and cognitive slowing: some convergent and divergent findings. *Psychol. Bull.* 128, 151–198. doi: 10.1037/0033-2909.128.1.151

Bertinetto, P. M., Burani, C., Laudanna, A., Marconi, C., Ratti, D., Rolando, C., et al. (2005). *CoLFIS (Corpus e Lessico di Frequenza dell'Italiano Scritto) [Corpus and Frequency Lexicon of Written Italian]*. Available online at: http://www.istc.cnr.it/grouppage/databases

Breitmeyer, B. G. (1984). *Visual Masking: An Integrative Approach*. New York, NY: Oxford University Press.

Burani, C., Thornton, A. M., and Zoccolotti, P. (2014). "*Literacy acquisition in Italian*," in *Reading Acquisition Across Languages and Writing Systems: An International Handbook*, eds L. Verhoeven and C. Perfetti (Cambridge, UK: Cambridge University Press).

Carver, R. P. (1992). Reading rate: theory, research, and practical implications. *J. Read.* 36, 84–95. doi: 10.1016/j.chb.2009.10.015

Cerella, J. (1985). Information processing rates in the elderly. *Psychol. Bull.* 98, 67–83. doi: 10.1037/0033-2909.98.1.67

Cerella, J. (1990). "Aging and information processing rate," in *Handbook of the Psychology of Aging, 3rd Edn.*, eds J. E.Birren and K. W. Schaie (San Diego, CA: Academic Press), 201–221.

Cerella, J., Poon, L. W., and Williams, D. M. (1980). "Age and the complexity hypothesis," in *Aging in the 1980s: Psychological Issues*, ed *L. W. Poon (Washington, DC: American Psychological Association)*, 332–340.

Chateau, D., and Lupker, S. J. (2003). Strategic effects in word naming: examining the route-emphasis versus time-criterion accounts. *J. Exp. Psychol. Hum. Percept. Perform.* 29, 139–151. doi: 10.1037//0096-1523.29.1.139

Chung, S. T., Mansfield, J. S., and Legge, G. E. (1998). Psychophysics of reading. *XVIII. The effect of print size on reading speed in normal peripheral vision. Vision Res.* 38, 2949–2962. doi: 10.1016/S0042-6989(98)00072-8

Coltheart, M., Davelaar, E., Jonasson, J. F., and Besner, D. (1977). "Access to the internal lexicon," in *Attention and Performance VI*, ed S. *Dornic (Hillsdale, NJ: Erlbaum)*, 535–555.

Coltheart, M., and Rastle, K. (1994). Serial processing in reading aloud: evidence for dual-route models of reading. *J. Exp. Psychol. Hum. Percept. Perform.* 20, 1197–1211. doi: 10.1037/0096-1523.20.6.1197

Coltheart, M., Rastle, K., Perry, C., Langdon, R., and Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychol. Rev.* 108, 204–256. doi: 10.1037/0033-295X.108.1.204

De Cara, B., and Goswami, U. (2002). Similarity relations among spoken words: the special status of rimes in English. *Behav. Res. Methods Instrum. Comput.* 34, 416–423. doi: 10.3758/BF03195470

De Mauro, T. (1999). *Grande Dizionario Italiano Dell'uso. [Large Italian Dictionary of Use]*. Torino: UTET.

Ellis, N. C., and Hooper, A. M. (2001). It is easier to learn to read in Welsh than in English: effects of orthographic transparency demonstrated using frequency-matched cross-linguistic reading tests. *Appl. Psychol.* 22, 571–599. doi: 10.1017/S0142716401004052

Ellis, N. C., Natsume, M., Stavropoulou, K., Hoxhallari, L., van Daal, V. H. P., Polyzoe, N., et al. (2004). The effects of orthographic depth of learning to read alphabetic, syllabic, and logographic scripts. *Read. Res. Q.* 39, 438–446. doi: 10.1598/RRQ.39.4.5

Enns, J. T., and Di Lollo, V. (2000). What's new in visual masking. *Trends Cogn. Sci.* 4, 345–352. doi: 10.1016/S1364-6613(00)01520-5

Faust, M. E., Balota, D. A., Spieler, D. H., and Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: implications for group differences in response latency. *Psychol. Bull.* 125, 777–799. doi: 10.1037//0033-2909.125.6.777

Felsten, G., and Wasserman, G. S. (1980). Visual masking: mechanisms and theories. *Psychol. Bull.* 88, 329–353. doi: 10.1037/0033-2909.88.2.329

Ferrand, L., and New, B. (2003). Syllabic length effects in visual word recognition and naming. *Acta Psychol. (Amst.)* 113, 167–183. doi: 10.1016/S0001-6918(03)00031-3

Fine, E. M., Peli, E., and Reeves, A. (1997). Simulated cataract does not reduce the benefit of RSVP. *Vision Res.* 37, 2639–2647. doi: 10.1016/S0042-6989(97)00051-5

Fine, E. M., Rubin, G. S., Hazel, C. A., and Petre, K. L. (1999). Are the benefits of sentence context different in central and peripheral vision? *Optom. Vis. Sci.* 76, 764–769. doi: 10.1097/00006324-199911000-00025

Fisher, D. F. (1975). Reading and visual search. *Mem. Cognit.* 3, 188–196. doi: 10.3758/BF03212897

Frith, U., Wimmer, H., and Landerl, K. (1998). Differences in phonological recoding in German- and English-speaking children. *Sci. Stud. Read.* 2, 31–54. doi: 10.1207/s1532799xssr0201_2

Hale, S., and Jansen, J. (1994). Global processing-time coefficients characterize individual and group differences in cognitive speed. *Psychol. Sci.* 5, 384–389. doi: 10.1111/j.0956-7976.2004.00689.x

Kinoshita, S., and Lupker, S. J. (2002). Effects of filler type in naming: change in time criterion or attentional control of pathways? *Mem. Cognit.* 30, 1277–1287. doi: 10.3758/BF03213409

Kwon, M., and Legge, G. E. (2012). Spatial-frequency requirements for reading revisited. *Vision Res.* 62, 139–147. doi: 10.1016/j.visres.2012.03.025

Kwon, M., Legge, G. E., and Dubbels, B. R. (2007). Developmental changes in the visual span for reading. *Vision Res.* 47, 2889–2900. doi: 10.1016/j.visres.2007.08.002

Lacouture, Y., and Cousineau, D. (2008). How to use MATLAB to fit the ex-Gaussian and other probability functions to a distribution of response times. *Tutor. Quant. Methods Psychol.* 4, 35–45.

Latham, K., and Whitaker, D. (1996). A comparison of word recognition and reading performance in foveal and peripheral vision. *Vision Res.* 36, 2665–2674. doi: 10.1016/0042-6989(96)00022-3

Lee, H. W., Kwon, M., Legge, G. E., and Gefroh, J. J. (2010). Training improves reading speed in peripheral vision: is it due to attention? *J. Vis.* 10, 1–15. doi: 10.1167/10.6.18

Legge, G. E., Mansfield, J. S., and Chung, S. T. L. (2001). Psychophysics of reading: XX. *Linking letter recognition to reading speed in central and peripheral vision. Vision Res.* 41, 725–743. doi: 10.1016/S0042-6989(00)00295-9

Logan, G. D. (1992). Shapes of reaction-time distributions and shapes of learning curves: a test of the instance theory of automaticity. *J. Exp. Psychol. Learn.* 18, 883–914. doi: 10.1037/0278-7393.18.5.883

Lupker, S. J., Brown, P., and Colombo, L. (1997). Strategic control in a naming task: changing routes or changing deadlines? *J. Exp. Psychol. Learn.* 23, 570–590. doi: 10.1037//0278-7393.23.3.570

Martelli, M., De Luca, M., Lami, L., Pizzoli, C., Pontillo, M., Spinelli, D., et al. (2014). Bridging the gap between different measures of the reading speed deficit in developmental dyslexia. *Exp. Brain Res.* 232, 237–252. doi: 10.1007/s00221-013-3735-6

Martelli, M., Di Filippo, G., Spinelli, D., and Zoccolotti, P. (2009). Crowding, reading and developmental dyslexia. *J. Vis.* 9, 1–18. doi: 10.1167/9.4.14

Martino, M. G., Pappalardo, F., Re, A. M., Tressoldi, P. E., Lucangeli, D., and Cornoldi, C. (2011). La valutazione della dislessia nell'adulto [A evaluation of dyslexia in adult readers]. *Dislessia* 8, 119–134.

Mayr, U., and Kliegl, R. (1993). Sequential and coordinative complexity: agebased processing limitations in figural transformations. *J. Exp. Psychol. Learn.* 19, 1297–1320. doi: 10.1037/0278-7393.19.6.1297

Medler, D. A., and Binder, J. R. (2005). *MCWord: An On-Line Orthographic Database of the English Language.* Available online at: http://www.neuro.mcw.edu/mcword

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. doi: 10.1037/0882-7974.7.2.257

Myerson, J., Hale, S., Zheng, Y., Jenkins, L., and Widaman, K. F. (2003). The difference engine: a model of diversity in speeded cognition. *Psychon. Bull. Rev.* 10, 262–288. doi: 10.3758/BF03196491

Paizi, D., Burani, C., and Zoccolotti, P. (2010). List context effects in reading words and nonwords in Italian: can the word frequency effect be eliminated? *Eur. J. Cogn. Psychol.* 22, 1039–1065. doi: 10.1080/09541440903216492

Palmer, E. M., Horowitz, T. S., Torralba, A., and Wolfe, J. M. (2011). What are the shapes of response time distributions in visual search? *J. Exp. Psychol. Hum. Percept. Perform.* 37, 58–71. doi: 10.1037/a0020747

Patel, T. K., Snowling, M. J., and de Jong, P. F. (2004). A cross-linguistic comparison of children learning to read in English and Dutch. *J. Educ. Psychol.* 96, 785–797. doi: 10.1037/0022-0663.96.4.785

Paulesu, E., McCrory, E., Fazio, F., Menoncello, L., Brunswick, N., Cappa, S. F., et al. (2000). A cultural effect on brain function. *Nat. Neurosci.* 3, 91–96. doi: 10.1038/71163

Pelli, D. G., Burns, C. W., Farell, B., and Moore-Page, D. C. (2006). Feature detection and letter identification. *Vision Res.* 46, 4646–4674. doi: 10.1016/j.visres.2006.04.023

Pelli, D. G., and Tillman, K. A. (2007). Parts, wholes, and context in reading: a triple dissociation. *PLoS ONE* 2:e680. doi: 10.1371/journal.pone.0000680

Pelli, D. G., Tillman, K. A., Freeman, J., Su, M., Berger, T. D., and Majaj, N. J. (2007). Crowding and eccentricity determine reading rate. *J. Vis.* 7, 1–36. doi: 10.1167/7.2.20

Perry, C., Ziegler, J. C., and Zorzi, M. (2007). Nested incremental modelling in the development of computational theories: the CDP+ model of reading aloud. *Psychol. Rev.* 114, 273–315. doi: 10.1037/0033-295X.114.2.273

Perry, C., Ziegler, J. C., and Zorzi, M. (2010). Beyond single syllables: large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. *Cogn. Psycol.* 61, 106–151. doi: 10.1016/j.cogpsych.2010.04.001

Potter, M. C. (1984). "Rapid serial visual presentation (RSVP): a method for studying language processing," in *New Methods in Reading Comprehension Research*, eds D. E. Kieras and M. A. Just (Hillsdale, NJ: Erlbaum), 91–118.

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychol. Bull.* 86, 446–461. doi: 10.1037/0033-2909.86.3.446

Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: fitting real data and failing to fit fake but plausible data. *Psychon. Bull. Rev.* 9, 278–291. doi: 10.3758/BF03196283

Ratcliff, R., Gomez, P., and McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychol. Rev.* 111, 159–182. doi: 10.1037/0033-295X.111.1.159

Raven, J. (2008). *SPM Standard Progressive Matrices*. Florence: Giunti O.S.

Rubin, G. S., and Turano, K. (1992). Reading without saccadic eye movements. *Vision Res.* 32, 895–902. doi: 10.1016/0042-6989(92)90032-E

Salthouse, T. A. (1985). *A Theory of Cognitive Aging.* Amsterdam: North-Holland.

Schmiedek, F., Oberauer, K., Wilhelm, O., Suß, H. M., and Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *J. Exp. Psychol Gen.* 136, 414–429. doi: 10.1037/0096-3445.136.3.414

Seidenberg, M. S., and McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychol. Rev.* 96, 523–568. doi: 10.1037/0033-295X.96.4.523

Seymour, P. H., Aro, M., and Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *Br. J. Psychol.* 94, 143–174. doi: 10.1348/000712603321661859

Strasburger, H., Rentschler, I., and Jüttner, M. (2011). Peripheral vision and pattern recognition: a review. *J. Vis.* 11:13. doi: 10.1167/11.5.13

Taylor, S. E. (1965). Eye movements in reading: facts and fallacies. *Am. Educ. Res. J.* 2, 187–202.

Taylor, T. E., and Lupker, S. J. (2001). Sequential effects in naming: a time-criterion account. *J. Exp. Psychol. Learn.* 27, 117–138. doi: 10.1037/0278-7393.27.1.117

Torgesen, J. K., Wagner, R. K., and Rashotte, C. A. (1999). *Test of Word Reading Efficiency*. Austin, TX: Pro-Ed Inc.

Tse, C. S., Balota, D. A., Yap, M. J., Duchek, J. M., and McCabe, D. P. (2010). Effects of healthy aging and early-stage dementia of the Alzheimer's type on components of response time distributions in three attention tasks. *Neuropsychology* 24, 300–315. doi: 10.1037/a0018274

Wagenmakers, E. J., and Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychol. Rev.* 114, 830–841. doi: 10.1037/0033-295X.114.3.830

Wagenmakers, E.-J., Grasman, R. P. P. P., and Molenaar, P. C. M. (2005). On the relation between the mean and the variance of a diffusion model response time distribution. *J. Math. Psychol.* 49, 195–204. doi: 10.1016/j.jmp.2005.02.003

Watson, A. B., and Pelli, D. G. (1983). QUEST: a Bayesian adaptive psychometric method. *Percept. Psychophys.* 33, 113–120. doi: 10.3758/BF03202828

Wilson, M. (1988). MRC psycholinguistic database: machine-usable dictionary, version 2.00. *Behav. Res. Methods Instrum. Comput.* 20, 6–11. doi: 10.3758/BF03202594

Wyse, D., and Goswami, U. (2008). Synthetic phonics and the teaching of reading. *Br. Educ. Res. J.* 34, 691–710. doi: 10.1080/01411920802268912

Yager, D., Aquilante, K., and Plass, R. (1998). High and low luminance letters, acuity reserve, and font effects on reading speed. *Vision Res.* 38, 2527–2531. doi: 10.1016/j.jcjo.2012.09.017

Yap, M. J., Balota, D. A., Sibley, D. E., and Ratcliff, R. (2012). Individual differences in visual word recognition: insights from the english lexicon project. *J. Exp. Psychol. Hum. Percept. Perform.* 38, 53–79. doi: 10.1037/a0024177

Yu, D., Cheung, S. H., Legge, G. E., and Chung, S. T. (2007). Effect of letter spacing on visual span and reading speed. *J. Vis.* 7, 1–10. doi: 10.1167/7.2.2

Yu, D., Park, H., Gerold, D., and Legge, G. E. (2010). Comparing reading speed for horizontal and vertical English text. *J. Vis.* 10, 1–17. doi: 10.1167/10.2.21

Zheng, Y., Myerson, J., and Hale, S. (2000). Age and individual differences in visuospatial processing speed: testing the magnification hypothesis. *Psychon. Bull. Rev.* 7, 113–120. doi: 10.3758/BF03210729

Ziegler, J. C., and Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory. *Psychol. Bull.* 131, 3–29. doi: 10.1037/0033-2909.131.1.3

Ziegler, J. C., Perry, C., Jacobs, A. M., and Braun, M. (2001). Identical words are read differently in different languages. *Psychol. Sci.* 12, 379–384. doi: 10.1111/1467-9280.00370

Zoccolotti, P., De Luca, M., Di Filippo, G., Judica, A., and Martelli, M. (2009). Reading development in an orthographically regular language: effects of length, frequency, lexicality and global processing ability. *Read. Writ.* 22, 1053–1079. doi: 10.1007/s11145-008-9144-8

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.

Received: 19 March 2014; accepted: 29 July 2014; published online: 19 August 2014.
Citation: Marinelli CV, Horne JK, McGeown SP, Zoccolotti P and Martelli M (2014) Does the mean adequately represent reading performance? Evidence from a cross-linguistic study. Front. Psychol. ***5****:903. doi: 10.3389/fpsyg.2014.00903*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

REVIEW ARTICLE
published: 15 September 2014
doi: 10.3389/fpsyg.2014.01023

List context effects in languages with opaque and transparent orthographies: a challenge for models of reading

Daniela Traficante [1,2] **** and Cristina Burani*** [3,4]

[1] *Department of Psychology, Catholic University of Milan, Milan, Italy*
[2] *NeuroMI, Milan Center for Neuroscience, Milan, Italy*
[3] *Institute for Cognitive Sciences and Technologies, ISTC-CNR, Rome, Italy*
[4] *Department of Life Sciences, University of Trieste, Trieste, Italy*

Edited by:
Davide Crepaldi, University of Milano-Bicocca, Italy

Reviewed by:
Petar Milin, Eberhard Karls University of Tübingen, Germany
Remo Job, University of Trento, Italy

****Correspondence:***
Daniela Traficante, Department of Psychology, Catholic University of Milan, Largo Gemelli 1, Milan 20123, Italy
e-mail: daniela.traficante@unicatt.it

This paper offers a review of data which show that reading is a flexible and dynamic process and that readers can exert strategic control over it. Two main hypotheses on the control of reading processes have been suggested: the route de-emphasis hypothesis and the time-criterion hypothesis. According to the former, the presence of irregular words in the list might lead to an attenuation of the non-lexical process, while the presence of non-words could trigger a de-emphasis of the lexical route. An alternative account is proposed by the time-criterion hypothesis whereby the reader sets a flexible deadline to initiate the response. According to the latter view, it is the average pronunciation difficulty of the items in the block that modulates the time-criterion for response. However, it is worth noting that the list composition has been shown to exert different effects in transparent compared to opaque orthographies, as the consistency of spelling-sound correspondences can influence the processing costs of the non-lexical pathway. In transparent orthographies, the non-lexical route is not resource demanding and can successfully contribute to the pronunciation of regular words, thus its de-emphasis could not be as useful/necessary as in opaque orthographies. The complex patterns of results from the literature on list context effects are a challenge for computational models of reading which face the problem of simulating strategic control over reading processes. Different proposals suggest a modification of parameter setting in the non-lexical route or the implementation of a new module aimed at focusing attention on the output of the more convenient pathway. Simulation data and an assessment of the models' fit to the behavioral results are presented and discussed to shed light on the role of the cognitive system when reading aloud.

Keywords: reading aloud, list context effects, models of reading, strategic behavior, orthographic systems

INTRODUCTION

During the last decades, since the pioneeristic work of Coltheart (1978), several studies on word recognition have found that changes in the stimuli list context can influence latency and accuracy in different tasks. These results challenge the assumption that word recognition is an automatic process for skilled readers (Underwood, 1978); in contrast, they suggest that strategic components can alter word processing in relation to the composition of the list context. Moreover, data from different languages have revealed a complex pattern of results and suggested that the characteristics of the language system, in particular its orthography-to-phonology consistency, could be considered as a "macro-context" in which the system may develop its specific setting, with potential consequences on the suitability of different strategies in different languages.

The most widely accepted reading models offer a framework to simulate the processes involved in the recognition of a single item, but do not consider the list context in which that item is presented. This review is aimed at showing that the data on list context effects call for a new approach in reading modeling, in which additional components and/or mechanisms are to be included to take into account strategic behavior.

After a brief description of the dual-route cascaded model (DRC), of the parallel-distributed-processing model (PDP), and of the connectionist dual-process model (CDP), empirical data drawn from different languages will be presented in order to highlight the role that list context and language context can play in implementing different strategies when reading aloud.

The large number of experiments assessing strategic effects in different tasks, such as lexical decision or semantic categorization, are not considered in the present paper for two main reasons. Firstly, we aim at providing evidence for the activation of strategic behavior in one task, reading aloud, in which decision-level processes are not assumed to be involved. Thus, we intend to avoid possible confounds between strategies triggered by the list context composition and decisional strategies that are operating in tasks such as lexical decision or sematic categorization. Secondly, only one reading model (Harm and Seidenberg, 2004) implements semantic components, due to the

high complexity of the model architecture required to take into account semantics. Accordingly, we thought it was more appropriate to consider only reading aloud studies, whose results can be simulated by means of the orthography-to-phonology mappings actually implemented by all the main computational models.

How data on list context effects may challenge the different modeling proposals and open new perspectives on the role of strategic control in reading aloud will be discussed in the final part of the paper.

FROM PRINT TO SOUND: MODELS OF READING AND BENCHMARK EFFECTS

The dual-route cascaded (DRC) model (Coltheart and Rastle, 1994; Coltheart et al., 2001) can be considered a computational evolution of the modeling tradition grounded in the 19th century modular approach. Despite its name, the model actually consists of three routes: the lexical semantic route, the lexical non-semantic route, and the grapheme-phoneme conversion (GPC) route (non-lexical route). However, the lexical semantic route has not been implemented yet (**Figure 1**). The model is cascaded because the activation is fed forward from one module to the following as soon as a process in that module starts, without waiting for the completion of the process itself.

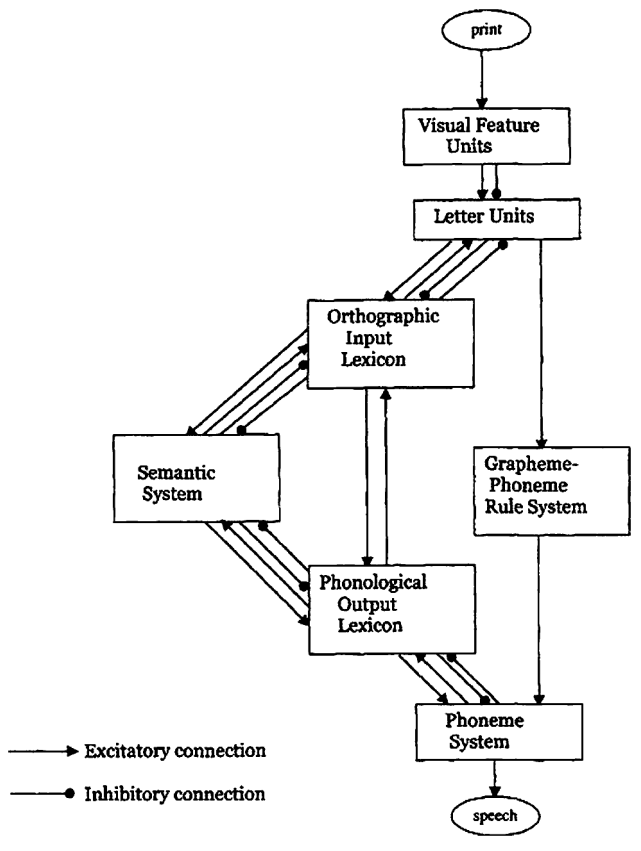


FIGURE 1 | The dual-route cascaded model. From "DRC: a dual route cascaded model of visual word recognition and reading aloud" by Coltheart et al. (2001), *Psychol. Rev.* 108, p. 214. Copyright 2001 by the American Psychological Association.

The early modules from print to word recognition (visual feature units, letter units, orthographic input lexicon) form a three-layer network, working with interactive activation and inhibition among the layers. In the case of non-words, no lexical entry can be addressed, but it is possible to produce a phonological output through the grapheme-to-phoneme correspondence (GPC) route. This route starts operating after a series of cycles from the input onset and converts letters to phonemes from left to right, serially, according to rules set on statistical grounds (Rastle and Coltheart, 1999). The generated phonemes add activation to units in the phoneme system, a layer common to both the lexical and non-lexical routes, in order to produce letter string pronunciation. However, non-words are not only read through the non-lexical route because they partially activate word neighbors[1] in the orthographic lexicon and these word units feed-forward activation to the phonological representations and to the phoneme system.

The need for implementing two different routes to read words and non-words has been challenged by the parallel-distributed-processing (PDP) model (Plaut et al., 1996). This is a one-route model of reading aloud, whose architecture is a three-layer network trained by an error-minimization learning algorithm. In the PDP model, all letter strings (both words and non-words) activate phonemic units in parallel. The distributional features of the input corpus are represented in the activation patterns within and between orthographical and phonological layers and all spelling-sound mappings depend on the parameter setting in the intermediate layer (hidden units). In this architecture, there are no specific pathways for reading words and non-words: "The information concerning spelling-sound correspondences, derived from exposure to actual words and encoded by the weights in such networks, is also used in generating pronunciations for unfamiliar stimuli" (Seidenberg et al., 1994, p. 1178).

The connectionist dual-process (CDP) model developed by Zorzi and colleagues (CDP: Zorzi et al., 1998; CDP+: Perry et al., 2007; CDP++: Perry et al., 2010) builds on the existing PDP and DRC models by combining features of both and is aimed at overcoming their limits. In the CDP model (**Figure 2**), spelling-sound connections are implemented, in parallel, *via* two pathways: a print-to-sound mapping mediated by lexical representations, implemented through a localist lexical route based on the interactive activation model as in Coltheart et al. (2001); a direct mapping from graphemic to phonemic units, implemented through a connectionist network (TLA: two-layer assembly model) as in Zorzi et al. (1998). This choice allows the CDP model to have not only an efficient solution to simulate lexical access in word reading, as in the DRC, but also a network for assembled phonology, that overcomes the absence of a learning mechanism in the DRC, a model which is fully hard-wired and whose non-lexical route works according to partially

[1]Orthographic neighbors of a string of letters have been operationalized by Coltheart et al. (1977) as the words that can be obtained by changing one letter and preserving the positions of the other letters. For example, neighbors of WORD are LORD, WARD, WORK.

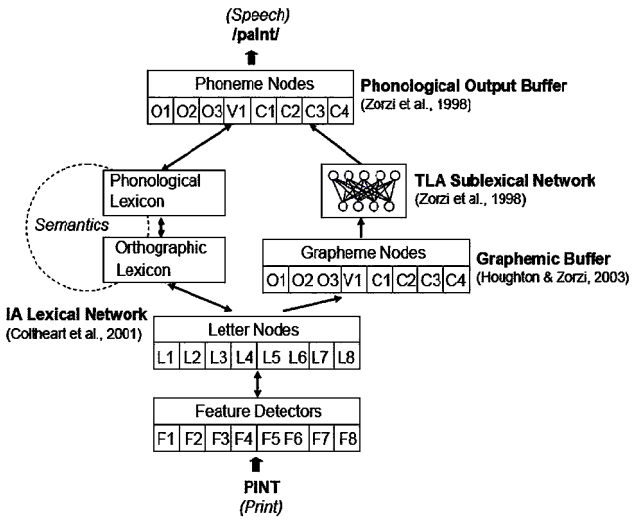


FIGURE 2 | The connectionist dual process model (CDP+). O = onset; V = vowel; C = coda; TLA = two-layer assembly; IA = interactive activation; L = letter; F = feature. From "Nested incremental modeling in the development of computational theories: the CDP+ model of reading aloud" by Perry et al. (2007), *Psychol. Rev.* 114, p. 280. Copyright 2007 by the American Psychological Association.

hand-coded sets of grapheme-to-phoneme conversion rules. Due to this network, the CDP model is able to simulate reading acquisition and developmental reading disorders, similar to the PDP model.

Pritchard et al. (2012) tested the DRC and CDP models in reading non-words comparing their performances to human responses. The DRC model showed a better match to participants' pronunciations (matching rates: 73.5% for DRC vs. 37.6% for CDP++). However, unlike behavioral data, the DRC model did not produce any lexicalization, while the CDP++ model produced very high lexicalization rates. In a recent paper, Perry et al. (2014) assessed the fit of DRC and CDP++ to the behavioral data in French, an orthography in which there are silent consonants at the end of words. The authors found that human readers, in reading non-words, tended to pronounce silent consonants that are not phonologically transcoded when detected in words. The DRC model, with the implementation of grapheme-to-phoneme rules for French, produced the pronunciation of these consonants in only 5.8% of the trials, while human readers pronounced them in 57.8% of the trials. The CDP++ model reached a rate of 41.2% pronunciations of silent consonants and this result was obtained through "a sublexical plus lexical analogy mechanism" (Perry et al., 2014).

Computational models test their claims to adequacy by simulating basic phenomena observed in reading aloud, that can be considered *benchmark effects* (Coltheart et al., 2001). In the present work only the benchmark effects, which have been proved to be influenced by stimulus list context will be presented: the lexicality effect, the length and length *by* lexicality effects, the word frequency effect, the regularity and regularity *by* frequency effects.

The *lexicality effect* (i.e., the observation that reading words is faster than reading non-words), in a lexicon-and-rule model like the DRC, is referred to the activation of different routes and modules, in relation to the lexical status of the stimulus. A non-word like BONT can activate some orthographic neighbors like FONT, BENT, BOND in the lexicon and gain activation in the phonemic system from these neighbors, but can be pronounced correctly only through a sequential activation of the phonemes corresponding to the graphemes B-O-N-T.

On the other hand, a regular item like WORD is likely to gain activation in the phonemic system from both routes, as both its lexical representation and the GPC rules produce a coherent phonemic pattern, which leads to a fast and correct response (Yates, 2010). In DRC and in CDP models, direct access to lexical representations, triggered by words, is faster than the serial application of GPC rules adopted in reading non-words, and this mechanism can explain why words are read faster than non-words. The PDP model offers an explanation in terms of frequency of activation of phonological patterns involved in the pronunciation of the target, assuming that non-words activate more rare orthographic-phonemic associations than words.

The serial processing of grapheme-to-phoneme conversion through the non-lexical route is also considered the mechanism that gives rise to the *length effect* (i.e., the longer the string of letters, the slower the reading latency). Overall, this effect is strong in reading non-words, while it is not found consistently in word reading (*length by lexicality effect*). While dual-route models can explain these effects quite well, as they assume that the sequential procedures involved in the grapheme-to-phoneme mapping are more time-costly than the direct access to lexical representations, the length *by* lexicality effect is particularly challenging for the PDP model. In fact, this model provides an account for the additional motor programming required by longer strings, but offers no ground for expecting differences in the visuo-perceptual scanning of words and non-words.

The *word frequency effect* (i.e., high frequency words are read faster and more accurately than low frequency words) is considered evidence for the activation of representations (either localist or distributed) in the orthographic lexicon (or system). All models assume that the speed of this activation is a function of the frequency of use of the corresponding words in the written language. Thus, lexical representations of high frequency words are activated faster than lexical representations of low frequency words.

Some interesting effects have been observed in reading exception words, like PINT or YACHT. These words are usually read more slowly than regular words such as FOND (*regularity effect*), but this effect is reliable only for low-frequency words (*regularity by frequency effect*). This phenomenon has been interpreted by the dual-route models as the result of an interference, in the phonemic system, between the output of the phonological lexicon and that of the grapheme-to-phoneme mapping mechanisms, which are doomed to fail. The PDP model refers this effect to the low level of activation of the phonological patterns involved in the pronunciation of low frequency exception words, but this model has some difficulties in simulating the pronunciation of a few low-frequency irregular words (e.g., AISLE). For this reason, Seidenberg et al. (1994) proposed that low-frequency irregular words can be read

through the semantic system, the third component of the so-called triangular model, not implemented by Seidenberg and McClelland (1989), but implemented in the model of Harm and Seidenberg (2004).

The ability to simulate the above mentioned benchmark effects has been considered a validity test for the computational reading models. However, it is worth noting that these effects are not found consistently in the behavioral data, as they can be influenced by list composition. In fact, the presence in the stimuli list of either words and non-words mixed together (*mixed context*) or of only one type of stimuli (only words or non-words: *pure context*) can alter the size of those effects. Moreover, data from different language contexts offer a complex picture with inconsistent results.

In the following sections, a review of some seminal works will be presented that can be considered representative contributions to the debate on the mechanisms underlying reading aloud in different list and language contexts (**Table 1**). In the section beneath, studies focusing on the issue of "which" pathway is mostly involved in different list contexts will be described, and the so-called *route de-emphasis* and *time-criterion* hypotheses will be introduced. The role of the consistency of the orthography-to-phonology correspondence will be discussed in the subsequent section, in which data on list context effects from both opaque and transparent orthographies will be presented.

In the final section, findings from the literature on the role of stimulus quality and proportion of related primes and targets in modulating frequency effects will offer further suggestions on the relation between list context effects and "how" the reading processes unfold. Proposals for new computational approaches, based on dynamic adaptation to the context conditions and trial history, will be presented and discussed, as they are likely to become the framework for research on reading processes in the next future.

"WHICH" PATHWAY FROM PRINT TO SOUND? ROUTE DE-EMPHASIS vs. TIME-CRITERION HYPOTHESIS

THE ROUTE DE-EMPHASIS HYPOTHESIS

Monsell et al. (1992) tested the strategic dissociation of lexical and non-lexical routes in English-speaking readers, focusing their attention on latency and accuracy when reading aloud non-words and exception words with the former assumed to be processed through GPC rules, the latter through the lexical pathway. In their experiment, participants had to read exception words and non-words, either in pure or in mixed blocks. Within word blocks, the frequency of use was blocked too, as the items of each block were all high frequency words or low frequency words. They found that reading high-frequency (very familiar) exception words was delayed by the presence of non-words in the list (mixed block), in comparison to the latencies observed in pure lists; however, when the participants expected to see low-frequency (less common) exception words, reading was not delayed by the presence of non-words in the list. On the other hand, reading of non-words was delayed, in comparison to a pure list condition in which only non-words were presented, by the presence of low-frequency exception words and not by the presence of high-frequency exception words. The authors proposed an explanation of these effects grounded on the distributions of processing times for the lexical and the non-lexical processes (**Figure 3**).

They made the assumption that, in the case of pure lists, the distribution of processing times for non-words has a large overlap with the distribution of processing times for low-frequency exception words, while the overlap with the distribution of high-frequency exception words is smaller (**Figure 3**: top). The non-lexical process should be slowed down in the case of mixed lists with low-frequency words, because the reader has to ignore the non-lexical output to increase the probability of a correct pronunciation of the exception words (**Figure 3**: bottom). On the side of low-frequency exception words, though, the slowing down of non-word processing does not have much effect, as the two distributions significantly overlap in any case, both in pure and in mixed conditions. As for high-frequency exception words, the expectancy of all exception words, as in a pure list, slows down the non-lexical process as well, and leads to faster RTs than in a mixed list as the spread between the two distributions increases. Monsell et al. (1992) proposed a continuous integration model of reading suggesting that "a phonological description is built up incrementally using fragments of information transmitted asynchronously from both processes" (p. 464). When information coming from the two processes is congruent, as in the case of regular words, the articulation of the currently available phonological description begins faster than when it is conflicting, as in the case of exception words. In the latter case, skilled readers can apply selective inhibition of (or inattention to) the non-lexical route, with different effects on latency distributions for high- and low-frequency exception words, as described above.

The effects of the presence of exception words in the experimental list on reading performance have been challenged by Coltheart and Rastle (1994). The authors aimed at assessing the strategic control operated by readers on the use of the lexical and non-lexical routes, by inserting different types of filler items in the naming experiments. They assumed that the presence of non-words should favor the use of the non-lexical route, while high-frequency exception words are expected to favor the use of the lexical route. The regularity effect, interpreted as an interference of the non-lexical route on lexical processing in reading exception words, should be larger when fillers are non-words than when they are high-frequency exception words, as the latter condition should induce neglect of the non-lexical route. Coltheart and Rastle's (1994) study did not confirm this hypothesis. However, in a further study, using only exception words with the irregular phoneme in the first position, Rastle and Coltheart (1999) found the expected list context effect. This means that a general slowing of the non-lexical route is triggered only when the inconsistency between the lexical and non-lexical routes already occurs at the beginning of the process (e.g., for words like "chef"); however, if the irregularity is in the middle of the exception words (e.g., "glow"), their lexical representation is accessed before the sequential letter-by-letter activation could interfere with lexical processing (**Figure 4**). These data are consistent with Monsell et al.'s (1992) findings, and support the route de-emphasis hypothesis.

This hypothesis has been implemented by modifying the parameters controlling the activation of the non-lexical route. Rastle and Coltheart (1999) successfully simulated the delaying effect of exception word fillers on naming regular word and

Table 1 | Pathway control and time criterion setting in reading processes: evidence accounted for in the present review.

Language	Reference	Main results	Suggestions for modeling
English	Monsell et al. (1992)	*High-frequency exception words*: delayed in mixed list with non-words; *Low-frequency exception words*: not delayed in mixed list with non-words; *Non-words*: delayed in mixed list with low-frequency words. No effects with high-frequency exception words.	In mixed lists with exception words, non-lexical route is inhibited (*route de-emphasis hypothesis*)
	Rastle and Coltheart (1999)	Regularity effect size is reduced when fillers are high-frequency exception words with irregular phoneme in the first position.	In reading exception words, the inconsistency between lexical and non-lexical routes triggers a general slowing of non-lexical route (*route de-emphasis hypothesis*)
	Zevin and Balota (2000)	*Low-frequency exception words*: more regularization errors when primed by non-words than by exception words; *Non-words*: slowed-down when primed by low-exception words	A separate control mechanism which computes the conflict between lexical and non-lexical route, and can slow down the non-lexical route in the case of exception words (*route de-emphasis hypothesis*)
	Reynolds and Besner (2008)	*Exception-words*: switch costs arise when exception word follows non-word *Low-frequency regular words*: switch costs neither after high-frequency exception words nor after non-words.	Separate control mechanism that modulates reading process on the basis of previous trials (*exogenous control*)
	Lupker et al. (1997)	*High-frequency exception words*: delayed in mixed list with non-words; *Low-frequency exception words*: faster RTs (but lower accuracy) in mixed list with non-words than in pure list; *Non-words*: faster RTs (but lower accuracy) in mixed list with high-frequency exception words and regular words than in pure list; *Regular words*: faster RTs in pure than in mixed list	Readers would set a time criterion to start articulation, according to the difficulty of the stimuli. In mixed list the criterion is beyond the preferred responding point for the fast stimuli but prior to the preferred responding point for the slow stimuli (*time-criterion hypothesis*)
	Kinoshita and Lupker (2003)	*Regularity effect* was not affected by prime lexicality (either low-frequency exception words or non-words); *Frequency effect* was reduced when words follow fast non-words; *Lexicality effect* was reduced when prime were (both slow and fast) non-words	*Lexical checking* is applied after a phonological code has been produced, in order to assess the matching between the codes generated by the two routes. The massive presence of non-words would lead to skip this check
Persian	Baluch and Besner (1991)	Transparent words: frequency and semantic priming effects only in pure context.	Decision mechanisms which selects the output from the routine (lexical or non-lexical) that first makes a response viable. In the case of the presence of non-words, only the non-lexical route is considered
Italian and English	Tabossi and Laghi (1992)	Italian: semantic priming in naming words only in pure list and not in mixed list. English: semantic priming in both pure and mixed context	In a *transparent orthography* the use of either the routes in reading words is influenced by the list context; in an *opaque orthography*, the use of lexical route in reading words is mandatory
Turkish	Raman et al. (2004)	Using non-words matched to words in reading speed, the *frequency effect* is always reliable	Also in a *transparent orthography* words are read through the lexical route, even though in mixed context with non-words Data on latency support the *time-criterion hypothesis*
Italian	Pagliuca et al. (2008)	*Lexicality effect* is reliable in both pure and mixed context; *Words* are read faster in the pure than in the mixed condition; *Non-words* are not influenced by the list context	Lexical route is never completely shut down, but is the main route to read words also in a transparent orthography; Data on non-words do not support the time-criterion hypothesis
Italian	Paizi et al. (2010)	Words: *frequency effect* reliable in both pure and mixed context; *Length effect* reliable only in mixed condition; Non-words: *length effect* reliable in all conditions	Data inconsistent with both de-emphasis and time-criterion hypotheses; Italian readers cannot block the activation of the lexical route, but have no reason for shutting down the non-lexical route, as it is not resource demanding in a transparent orthography like Italian

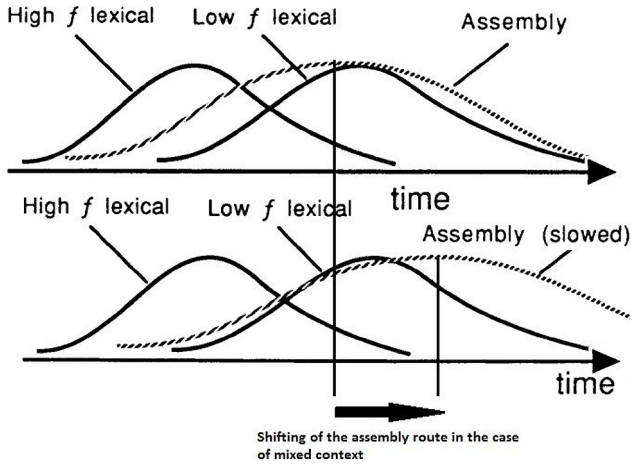


FIGURE 3 | Top: imaginary distributions of processing time for high- and low-frequency exception words (*High f lexical, Low f lexical,* respectively; solid curves) and for non-words (sublexical *assembly* process; broken curve). **Bottom:** the same, with the distribution for the assembly process shifted to the right, to simulate the hypothesized effect of trying to ignore assembled output. Adapted from "Lexical and sublexical translation of spelling to sound: strategic anticipation of lexical status" by Monsell et al. (1992), *J. Exp. Psychol. Learn. Mem. Cogn.* 18, p. 463. Copyright 1992 by the American Psychological Association.

non-word targets by increasing the number of cycles (from 17 to 22) elapsed before the non-lexical route can process the next letter. Perry et al. (2007) proposed a similar parameter manipulation, increasing the number of cycles occurring between the processing of each letter in the non-lexical pathway of CDP+ from 15 to 17. They obtained a delay in reading non-words (7.94 cycles) and words (2.94 cycles) proportional to the delay in behavioral data (20 and 12 ms, respectively: Rastle and Coltheart, 1999, p. 494, Table 5). In contrast, the modification to the DRC made by Rastle and Coltheart (1999) led to an overestimation of the delay for non-words (22.49 cycles) and an underestimation for words (0.56 cycles; p. 495, Table 6).

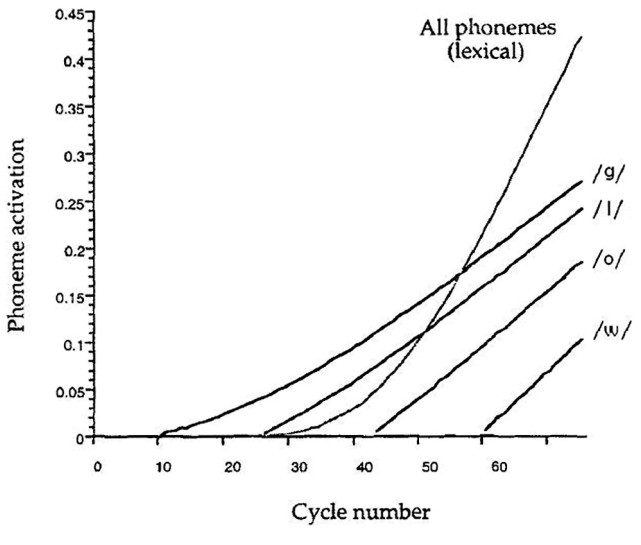


FIGURE 4 | Lexical and non-lexical activation of the phonemes of the exception word GLOW during its reading by the DRC model. Adapted from "DRC: a dual route cascaded model of visual word recognition and reading aloud" by Coltheart et al. (2001), *Psychol. Rev.* 108, p. 234. Copyright 2001 by the American Psychological Association.

Zevin and Balota (2000), in order to account for the dependence of reading on specific sources of information, used a priming procedure in which each trial consisted of five primes followed by a target and all the stimuli had to be read aloud. This procedure was aimed at creating "a situation in which dependence on the most efficient pathway for processing the prime stimuli would be maximally beneficial" (p. 123). Their results showed that non-word naming is slowed down after naming a sequence of (five) low-frequency exception word primes – a condition in which the non-lexical route is de-emphasized. Moreover, they found that low-frequency exception words gave rise to more regularization errors when primed by a sequence of (five) non-words than by other exception words, as the route de-emphasis hypothesis predicts.

In order to simulate the route de-emphasis effects, Zevin and Balota (2000) proposed adding a separate control mechanism to the DRC model which obtains information from the phoneme system and computes "the conflict or the ratio of contributions between the lexical and sublexical routes on a given trial" (Zevin and Balota, 2000; p. 132). This control system can slow down the sublexical route, as proposed by Rastle and Coltheart (1999) in the case of exception word primes, or change strategy, by gaining from the outputs of both routes, in the case of non-word primes, because "... readers are sensitive to the processing demands presented by different stimuli in a word-naming task and [...] they are able to adjust their dependence on different sources of information accordingly" (Zevin and Balota, 2000; p. 133). The same mechanisms might in principle apply also to a PDP model to adjust the relative contribution of the direct orthography-to-phonology network and of the semantic system in spelling-to-sound translation.

However, it is an open question whether the changes in strategy are triggered by the features of the item itself (*exogenous control*) or by the reader's expectations (*endogenous control*), developed on the basis of the trial sequence. Some suggestions on this issue come from studies carried out with the task-switching paradigm. Reynolds and Besner (2008), in studying the use of the reading routes, adopted the alternating runs paradigm which consists of presenting participants with two tasks in a predictable AABB sequence. With this paradigm, RTs on switch trials (A→B) are usually slower than on stay trials (A→A). The switch costs are interpreted as the output of an exogenous control component of the process, driven by the presentation of the task-relevant stimulus.

For instance, when the A task consists of reading high-frequency exception words and the B task in reading non-words (Reynolds and Besner, 2008: Experiments 1–3), switch costs arise from the interference between the lexical strategy and the non-lexical strategy, caused by the presentation of a non-word. The same happens when an exception word follows a non-word. Switch costs were neither found when low-frequency regular words were presented after high-frequency exception words (Experiment 4)

nor when the same regular words were in an alternating sequence with non-words (Experiment 5). These results seem to confirm that low-frequency words are likely to be read through the lexical route when mixed with high-frequency exception words and through the non-lexical route when presented with non-words. The switch costs are consistent with the proposal of a separate control mechanism that modulates reading process on the basis of previous trials.

THE TIME-CRITERION HYPOTHESIS

A different perspective has been offered by Lupker et al. (1997), Kinoshita and Lupker (2002, 2003), Chateau and Lupker (2003). In keeping with Monsell et al.'s (1992) results, Lupker et al. (1997) found that, in English, both high-frequency exception words and regular words were read faster in a pure than in a mixed condition. But, in contrast with Monsell et al.'s (1992) results, non-words were named faster when mixed with words (both high-frequency exception words and regular words) than in a blocked condition. Moreover, low-frequency exception words achieved faster RTs in a mixed condition with non-words. The data on low-frequency words and on non-words, however, showed that the gain in RTs occurs at the expense of accuracy, as there was a trade-off between latencies and accuracy for these stimuli.

In order to interpret their results, Lupker et al. (1997) proposed the *time-criterion hypothesis*: readers would set a time criterion to start articulation, which is determined by the difficulty of the stimuli and is aimed at maintaining an acceptable level of accuracy and rapidity. In this way, when easy (regular words, high-frequency exception words) and difficult (non-words, low-frequency exception words) stimuli are mixed together, "the criterion would have tended to stabilize at a point that was beyond the preferred responding point for the fast stimuli but prior to the preferred responding point for the slow stimuli" (Lupker et al., 1997; p. 578). This claim should also explain the trade-off between latencies and errors: the early start of the articulation of a difficult stimulus can lead to a lower level of accuracy. According to the authors, the criterion can also be set in relation to the task and to the experimental procedure, since stressing speed or accuracy can produce different effects.

Kinoshita and Lupker (2003) adopted the priming paradigm introduced by Zevin and Balota (2000), but selected three types of primes: fast non-words (short and with high N-size), slow non-words (long and with low N-size) and low frequency exception words. They studied the influence of prime lexicality (words/non-words) and of prime type (slow vs. fast stimuli) on the size of three benchmark effects, namely the regularity effect, the frequency effect and the lexicality effect. The regularity effect was significant and it was not affected by the prime lexicality (words/non-words). However, the reading latencies of the targets were affected by the prime type, as targets were named faster when following faster primes, irrespective of being exception words or (fast) non-words (Experiment 1). The frequency effect was reduced in the case of a fast non-word context, but not in the case of a slow non-word context. The authors interpreted this result as evidence that a context composed of rapidly named stimuli reduces the difference between high and low frequency words, because of a floor effect for high frequency words (Experiment 2). While the results of the first two experiments were in line with the time-criterion hypothesis, the third experiment showed inconsistent data. In fact, the lexicality effect was reduced in the case of non-word (both fast and slow) primes. The authors interpreted this result as a consequence of the application of a second – lexical checking – reading strategy, according to which "prior to emitting a naming response, readers have the option of consulting the phonological output lexicon in order to determine whether the code generated by the phonological coding process matches a code in the output lexicon" (Kinoshita and Lupker, 2003; p. 412). Lexical checking would take place after a phonological code has been produced and can be skipped in the case of a massive presence of non-word stimuli.

The complex pattern of results described above offers a view of reading as a dynamic process, in which different procedures for obtaining phonology from print (the lexical or non-lexical pathways) can be strategically activated according to the characteristics of the list context. In the following section, studies will be reported aimed at assessing whether the consistency of grapheme-to-phoneme correspondence may introduce further differences among languages in the way in which strategic control is applied. In fact, a fully consistent orthography could make it possible, in principle, to read words through the only involvement of the non-lexical route and this opportunity might give rise to a completely different pattern of list context effects, in comparison to opaque orthographies. Data from neuroimaging studies support the view of different reading processes in English, that has an opaque orthography, as opposed to Italian, whose orthography is considered to be transparent (Paulesu et al., 2000). Also the psycholinguistic grain size theory (Ziegler and Goswami, 2005, 2006; Goswami and Ziegler, 2006) pointed out that children of transparent orthographies would learn reading by relying on small units of phonological recoding, while children of opaque orthographies are supposed to use multiple phonological recoding strategies, based on larger units, to avoid mispronunciations. It can thus be assumed that differences in the early phases of learning to read might produce different reading behaviors in the mature system.

EVIDENCE FROM DIFFERENT LANGUAGES: THE ROLE OF ORTHOGRAPHIC CONSISTENCY

To shed light on the two main perspectives described above (de-emphasis hypothesis and time-criterion hypothesis), in the context of a transparent orthography like Turkish, Raman et al. (2004) followed Kinoshita and Lupker's (2003) approach. They manipulated non-word length, in order to create two lists of non-words that were matched on reading time (rather than length) to high-frequency and low-frequency words, respectively. Thus they obtained four different lists of stimuli: (a) high-frequency words and (b) corresponding fast non-words, (c) low-frequency words and (d) corresponding slow non-words. By using non-words matched to words in reading speed, the authors could test, separately, the effect of the time-criterion induced by the reading time of the stimulus context and the effect of shifting from the lexical to the non-lexical route in the presence of non-word stimuli. To assess the involvement of the lexical route, the authors analyzed the size of the word frequency effect in different list

contexts, assuming that, following the de-emphasis hypothesis, the frequency effect should not be reliable in mixed condition (words and non-words together), in spite of its reliability in pure (only words) condition.

Contrary to the latter prediction, Raman et al. (2004) found that the frequency effect was significant in all conditions, even though its size was modulated by list composition. The pattern of results shows that, also in a transparent orthography, words are read through the lexical route as the frequency effect is reliable in all conditions irrespective of the presence of non-words in the list. Moreover, the naming latencies for high-frequency words are influenced by the mean difficulty of the list thus supporting the time-criterion hypothesis. Accordingly, the authors claimed that "it appears that neither the lexical nor the non-lexical route is under strategic control of Turkish readers and that the data are best explained by a time criterion position" (Raman et al., 2004; p. 498).

This conclusion is not consistent with previous results in other languages with transparent orthographies, like Persian (Baluch and Besner, 1991) and Italian (Tabossi and Laghi, 1992). In fact, in those studies the presence of non-words eliminated lexical and semantic effects in reading words. In particular, Baluch and Besner (1991) found that Persian transparent words (i.e., with printed vowels) are named by using the non-lexical route (as indicated by the absence of semantic and frequency effects) when presented in a mixed context with non-words, while they are likely to be read by means of the lexical route (indicated by the presence of semantic and frequency effects) when presented in a pure context. To explain these results, the authors proposed a decision mechanism which selects the output from the route (lexical vs. non-lexical) that first makes a response available. In the case of the presence of non-words, such a mechanism would first consider the output of the non-lexical route, thus eliminating lexical and semantic effects in transparent word reading. In the case of words alone, only the lexical route would be selected. Raman et al. (2004) interpreted the lack of significance of frequency effect in the study by Baluch and Besner as due to the use of non-words matched to the words in length and not in reading speed. Such non-words are likely to be so difficult to lead the time criterion at a very high level, slowing down high-frequency words to such an extent that the frequency effect is eliminated.

However, Tabossi and Laghi (1992), for the Italian language, also came to a conclusion consistent with Baluch and Besner's (1991) results. They adopted the same experimental design to assess list context effects in two orthographies with different degrees of spelling-sound consistency: Italian and English. The authors found different list context effects for the two languages. For Italian, semantic priming in naming words occurred only in a pure context (words alone), while in a mixed context, in which both words and non-words were presented, semantic priming was not significant. These data suggest that when non-words are present, Italian words are likely to be read through the non-lexical route. In contrast, in English, semantic effects were found not only in pure, but also in mixed contexts. The results indicate that in reading a language with an opaque orthography lexical access is mandatory.

It is worth noting, however, that both in Baluch and Besner's (1991) and in Tabossi and Laghi's (1992) studies, some evidence for the activation of the lexical pathway in reading words was also found in a mixed context. In Persian, the lexicality effect emerged in all list contexts, with transparent words being read faster than non-words in both pure and mixed lists. This result shows that words, differently from non-words, can gain activation not only from the non-lexical route, but also from the phonological output lexicon, which feeds forward to the phonemic output buffer and makes word naming faster than non-word naming. Overall, this study shows the high flexibility of the word-naming process in a transparent orthography like Persian.

Similarly, Tabossi and Laghi (1992) demonstrated that it is possible to obtain semantic effects also in Italian, in a mixed context, by adding to the list of stimuli a small proportion (about 20%) of trisyllabic words stressed on the first syllable (e.g., *fàcile*, easy). In order to correctly name these words, Italian readers have to access lexical knowledge, while reading them through the non-lexical route is likely to lead to the default stress assignment (valid for about 70% of Italian words) on the penultimate syllable (e.g., *facìle**), that produces a wrong response. In this condition, the authors found effects of semantic priming in Italian, just as in English. These data led the authors to conclude that skilled readers rely on their lexical knowledge in naming most common words, regardless of the different writing systems. Only in unusual conditions, in which they read lists of non-words and regular words, readers of transparent orthographies can find it more useful to apply the non-lexical assembled phonology. Switching from lexical reading to the unusual non-lexical route is a matter of strategy that educated adults can apply even if they may be unaware of this.

Several years later, Pagliuca et al. (2008) came to similar conclusions. They tested list context effects on word and non-word reading in Italian, contrasting the de-emphasis and time-criterion hypotheses. They presented readers with high-frequency and low-frequency Italian words in pure and mixed conditions with non-words. Words in the pure condition were read faster than in the mixed condition and this evidence is consistent with both the route de-emphasis and the time-criterion hypothesis. In contrast, reading non-words was not influenced by the list context at all and this result cannot be accounted for by the time-criterion hypothesis. Furthermore, the authors found that, also in a transparent orthography like Italian, the lexicality effect is reliable in all conditions (pure and mixed), even when non-words are compared to low-frequency words. Pagliuca et al. (2008) concluded that "these data support the view that the lexical route is never completely shut down but is instead the main route used in naming words, regardless of orthography depth" (Pagliuca et al., 2008; p. 431).

Strong support for the use of the lexical pathway in a transparent orthography comes from further research conducted in Italian, in which the authors (Paizi et al., 2010) adopted an experimental design similar to Raman et al. (2004). The main difference was that non-words were not matched to words on reading times, but on length in letters, N-size, bigram frequency, orthographic rules, and initial phoneme. Paizi et al. (2010) also tested the effect of stimulus length, as the role of length in reading low-frequency words and

non-words could be ascribed to the use of the non-lexical route. They found that the frequency effect was reliable in all conditions, even when words were mixed with non-words. These data are not consistent with previous research that found for transparent-orthography languages the reduction (Raman et al., 2004) or the disappearance (Baluch and Besner, 1991; Tabossi and Laghi, 1992) of the frequency effect in mixed conditions. In contrast to the stability of the frequency effect in all list conditions, the effect of length for words was fully significant only in the all-mixed condition in which high frequency and low frequency words were presented mixed with each other and to their corresponding non-words. For non-words the effect of length was fully reliable in all conditions. Hence, Paizi et al.'s (2010) data do not support the route de-emphasis account, as the reliability of the frequency effect in all conditions calls for a constant involvement of lexical activation, irrespective of the presence of non-words in the list. However, these data do not support the time-criterion account either, as there are no relevant differences across conditions for any kind of stimuli. The authors proposed that for Italian readers it is impossible to block the activation of the lexical route (as suggested by the persistence of the frequency effect). However, Italian readers also have no reason for shutting down the non-lexical route when reading words (as indicated by the varying length effect for words in the presence of a constant length effect for non-words), because, due to the ease of applying rules of print-to-sound conversion, the non-lexical route in Italian is not resource demanding. This latter interpretation also applies to the absence of any influence of list context in non-word reading reported by Pagliuca et al. (2008).

OPEN QUESTIONS AND NEW APPROACHES: DOES LIST CONTEXT AFFECT "HOW" PROCESSING UNFOLDS?

The complexity of the results summarized above indicates that current theories and models are far from providing an adequate understanding of the mechanisms actually involved in reading processes. O'Malley and Besner (2008) claimed that one of the limits of the main computational approaches in this field is the assumption of cascaded activation as a fix processing mode. The authors suggested that the experimental context has an influence not only on *what/which* pathway is involved or slowed down, but also on *how* processing unfolds over time, i.e., whether the mechanisms that rule the functioning of the system may change and *why*, in case a change is triggered. In other words, they proposed that the list context may lead to modifications in the modality in which the decoding process is implemented and this change is detectable only by considering joint effects of different variables, which tap into different processing steps.

O'Malley and Besner (2008) offered evidence in favor of their view by jointly analyzing and modeling the effects of stimulus quality and frequency. They started from the observation that evidence from the lexical decision task shows additive effects of stimulus quality and frequency on RTs (Stanners et al., 1975; O'Malley et al., 2007; Yap and Balota, 2007), while data on reading aloud support interactive effects between the two variables (O'Malley et al., 2007; Yap and Balota, 2007). They demonstrated, in a set of three experiments, that the inconsistency between the results in the two tasks is not due to the task itself, but to the presence of non-words in the lexical decision procedure and their absence in the reading aloud experiments. In fact, asking participants to read aloud words and non-words in a mixed list, they obtained the same additive effect observed in lexical decision, while data from reading aloud a pure list of words showed interactive effects.

The authors, in the framework of the DRC model, advanced the *lexicalization hypothesis*, according to which, in reading a mixed list of words and non-words, when stimulus quality is low, the system would use a thresholded mode of processing at the letter level, in order to prevent lexicalization errors in reading non-words. This processing mode would stop the cascaded feed-forwarding of the activation from the letter level to the GPC module and to the orthographic lexicon, getting the system to work in a sequential way (Sternberg, 1969). As the stimulus quality lowers, the higher the threshold will be for activation of letter nodes. After that level, the process continues to operate in its usual mode, with parallel activation of lexical representations (for words) and sequential implementation of the GPC rules (for non-words). Thus the effects of stimulus quality and word frequency will be additive. In the case of pure lists of words, the threshold of the letter level is not required, as only lexical representations are involved, so a low level of stimulus quality would interfere more with the activation of low frequency than of high frequency representations. As a consequence, an overadditivity of the frequency effect would appear.

The CDP+ model (Ziegler et al., 2009) offers mechanisms useful to simulate the suggestions made by O'Malley and Besner (2008). In fact, in the CDP+ model, the non-lexical route reaches a threshold, while the lexical route is cascaded. Thus, the observation of additive effects would depend on the strength of the non-lexical route in comparison to the lexical one: in naming a mixed list of words and non-words the lexical route would be de-emphasized, in order to avoid lexicalization of non-words, and in this condition additive effects would appear. Moreover, in the case of a very low stimulus quality, the sensible reduction of the activation in the lexical route would lead to a small word frequency effect, giving rise to an underadditive effect, with high frequency words affected more by low stimulus quality than low frequency words.

Interesting clues for understanding how the context can influence word processing arise from a recent work by Scaltritti et al. (2013). In a naming task in which semantic priming, stimulus quality and frequency effects were assessed, the authors presented both words and non-words. They found an additive effect of stimulus quality and frequency in the case of related primes, but an overadditivity effect in the case of unrelated primes, since low-frequency words were more disrupted by low stimulus quality than high frequency words, as in Borowsky and Besner's (1993) study. The authors explained these results in terms of the *prime reliance account*, according to which the reliance on prime information is higher in the case of degraded stimuli than in the case of clear stimuli. The prime information is particularly helpful for low-frequency degraded words and this support can compensate for the disruptive effect of low stimulus quality, decreasing the likelihood of an interaction between stimulus quality and frequency. On the contrary, in the case of unrelated primes, the low-frequency degraded words cannot gain advantage from the prime, so they are

particularly disrupted in comparison to high-frequency words. In this case, an overadditivity effect is likely to emerge. These results show that the reliance on the prime can be considered a strategy influenced by the list composition: in the case of all unrelated prime-target pairs (Scaltritti et al., 2013: Experiment 2), the information from the prime is skipped and only the expected additivity effect of stimulus quality and frequency is found (see O'Malley and Besner, 2008).

According to the *episodic account* (Bodner and Masson, 2003), reliance on the prime can be varied according to the proportion of trials in which prime and target are semantically related (RP: *relatedness proportion*) and can produce different biases on RTs. If prime reliance is high (high RP), then related-prime targets would speed up, while unrelated-prime targets should be slowed down, due to the potential interference from the prime. Results from a lexical decision task with masked priming (Bodner and Masson, 2003) showed that semantic priming is higher when RP is 0.80 than when RP is 0.20, but no clear inhibition was observed for unrelated-prime targets and this result is inconsistent with an episodic account.

Bodner and Masson (2003), in a further experiment (Experiment 2), found a similar effect of RP, also in the case in which 80% of related primes had different relatedness with the target. In the experiment, high RP condition was made of 20% of semantically related prime-target pairs (e.g., nurse–DOCTOR), the same as low RP condition, and 60% of repetition primes (e.g., doctor–DOCTOR). The authors suggested that their results are neither consistent with the idea of automatic spreading activation within the orthographic lexicon, nor with consciously controlled processes like expectancy and semantic matching, because the prime-target SOA of 45 ms should not be long enough to carry out such processes. These data would suggest that "enhanced reliance on masked prime resources operates in a rather general manner, making use of whatever relation holds between a related prime and target" (p. 650) and indicate the role of prime reliance in creating "a form of episodic resource that can be recruited to assist with target processing" (p. 651).

Within the PDP approach, Plaut and Booth (2000, 2006) proposed a sigmoidal function relating input to output, that offers an interesting framework to interpret "how" the previous trials and/or the list context can influence the size of additivity or interaction effects. According to this function (see **Figure 5**), if the input activation values of all target types (e.g., the four points in a 2 × 2 factorial design: high-low frequency, degraded-clear stimulus quality) are in correspondence with the steep section of the curve, there is a quasi-linear increment of the effect size for both conditions, with an additive effect in the output values. If the value of one or more target types is associated with activations corresponding to different sections of the curve, then overadditive (highest part of the curve) or underadditive (lowest part of the curve) effects in output values are expected.

The model of Plaut and Booth (2000) offers an interesting account of different kinds of effects across variables, but it leaves open the question concerning the variables that can determine the change in the level of activation on the input axis. The work of Kinoshita et al. (2011) provides possible suggestions in underscoring the role of recent trials on the processing of a current

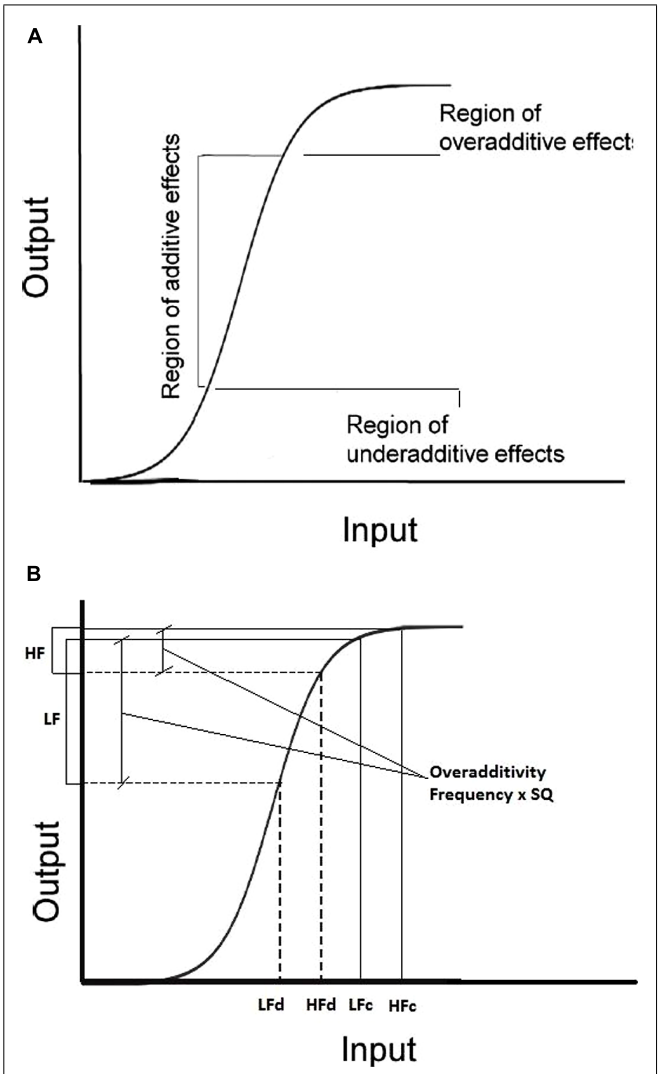


FIGURE 5 | Representation of the Plaut and Booth (2000) input–output activation function. (A) regions of the curve corresponding to underadditive, additive, and overadditive effect. **(B)** example of application of the function to data from a 2 × 2 factorial design Frequency (High-Low) × Stimulus Quality (SQ: degraded-clear).

stimulus. Their model, the adaptation to the statistics of the enviroment (ASE), is based on the results obtained from a linear mixed-effect model analysis (Baayen et al., 2008). This analysis allowed the authors to prove the effect of the previous trial on the processing of the target, both as a main effect and in interaction with other features of the context and of the current stimulus. This model mirrors the time-criterion hypothesis described above, as it assumes that after an easy stimulus, the latency in the next trial will decrease, while after a difficult stimulus, the latency will increase.

An interesting application of an integrated approach of the two models (Plaut and Booth, 2000; Kinoshita et al., 2011) can be found in Masson and Kliegl's (2013) work. In their experiments, Masson and Kliegl (2013) adopted a semantic priming paradigm, with a prime-target SOA of 200 ms, and varied the

stimulus quality. In the ANOVA on aggregated data, they found the usual additivity effect between frequency and stimulus quality, but analyses through the mixed-effect model revealed a completely different pattern of results. The following variables were entered in the model: priming relation (related–unrelated), word frequency (high–low) and stimulus quality (clear-degraded) of the target corresponding to the analyzed RT; the lexical status and the stimulus quality of the last-trial target. All three variables (priming relation, frequency, stimulus quality) characterizing the target were significant and showed a pattern of additivity as expected according to the literature and to the results found with the ANOVA technique.

Additionally, the variables referring to the last-trial target were involved in a significant interaction with the three target variables: if the last-trial target was a degraded non-word, priming was more effective for low-frequency targets and almost nil for high-frequency words, giving rise to an overadditivity effect. This result is consistent with the ASE model, because an extremely difficult item such as a degraded non-word is likely to require more evidence in responding to the next trial. If this increased activation in the input signal is represented in the sigmoid curve proposed by Plaut and Booth (2006), the highest part of the sigmoid curve is involved (see **Figure 5**), thus an overadditivity effect appears. On the contrary, when the last-trial target was a clear word, priming was effective only for high-frequency targets (underadditivity). The underadditive effect in the case of a clear previous word could be explained with the reverse reasoning: "a less demanding experience on trial $n - 1$ might allow the output activation threshold to be lowered, moving the criterion back down the sigmoid function" (p. 906). This shift would produce the observed underadditive effect.

These results proved that recent trial history can exert an important influence on word processing. Considering this component, Masson and Kliegl (2013) claim that the additivity effect between frequency and stimulus quality can also be described as the consequence of two opposite interactions: an overadditivity effect, when the input activation required for responding to the target is high, due to the difficulty of the last-trial target (e.g., degraded non-word), and an underadditivity effect when the last-trial target is easy (e.g., clear word) and the required input activation is low.

O'Malley and Besner (2013) observed that the effects found by Masson and Kliegl (2013) could be "a reflection of decision-level processes specific to the lexical decision task" (p. 1322), so they examined the effect of prior trial history in reading aloud tasks. They found a main effect of prior trial history, but no interactions between this factor and each of the two main factors – stimulus quality and word frequency – was observed. The authors ascribed the effects observed by Masson and Kliegl (2013) to the presence of semantic priming in the lexical decision task, that would promote retrospective processing in a significant way in comparison with other tasks, such as reading aloud.

However, even Masson and Kliegl (2013) failed to find the expected overadditivity effect when the last-trial target was a degraded non-word and the target was not primed. This condition, arguably the most difficult one, gave rise to an underadditivity effect when the stimulus quality was kept constant within a block of trials. The authors suggested that this anomalous outcome could be "the product of a ceiling effect on response time in the slowest condition (low-frequency and degraded target)" (p. 909), that would prevent appreciating a significant change in RTs in comparison to high-frequency words.

Overall, not even the combination of the ASE model (Kinoshita et al., 2011) with the activation function proposed by Plaut and Booth (2000) is able to thoroughly explain all behavioral data on list context effects, but it is a new and interesting approach, that offers some hints for modeling reading processes and for carrying on data analysis in experimental research.

CONCLUSION

Experimental evidence on list context effects reveals that pronouncing a string of letters is considerably more than an automatic process. In a very simple task like the naming of single items, there are complex interactions among the stimulus properties (psycholinguistic features and stimulus quality), the list context (pure/mixed block), and the properties of the previous stimulus in the list. In addition, data from several languages also show that the orthography-phonology consistency may have a role in determining the usefulness of different strategic settings of the system.

In opaque orthographies, several stimuli are likely to be read correctly only through the lexical pathway (e.g., exception words: PINT, YACHT, etc.), whereas in transparent orthographies most of the words can be read correctly through grapheme-to-phoneme conversion. Hence, skilled readers of opaque orthographies are more likely to be used to shutting down the non-lexical pathway than skilled readers of transparent orthographies. In fact, the non-lexical pathway in transparent orthographies is not very resource-demanding and skilled readers may use it in a highly efficient way. The efficiency in the use of the two pathways develops during literacy acquisition, as some studies on children with and without developmental dyslexia suggest (see Paizi et al., 2011).

How far do current computational models of reading account for the flexibility of the cognitive system and the results of the interaction between orthography and reading processes? The class of dual-route models could be considered more consistent with the de-emphasis hypothesis. In fact, in these models (DRC, CDP) the modification of parameter setting in the non-lexical pathway can be enough to implement list context effects. However, in this framework a new component is required, assumed to operate in two different ways: either choosing which route is to be de-emphasized, or deciding which of the two outputs (from the lexical and from the non-lexical pathway, respectively) is to be taken into account. Moreover, differences in the orthographic consistency of the language can influence the usefulness of de-emphasizing the lexical or the non-lexical route.

The time–criterion hypothesis offers an interpretation of list context effects that is independent of any specific pathway or control mechanism, while introducing the view of reading as a dynamic process, in which the overall level of activation is a function of previous trials. The ASE model, grounded on mixed-effect model statistics, is a recent formal description of the time-criterion hypothesis which, integrated with the Plaut and Booth's activation function, gives a flexible and probabilistic

framework for interpreting additivity and interaction effects. The PDP model, which includes learning mechanisms and is a one-route network, seems to be consistent with this hypothesis. However, in principle, the trial-to-trial changes originating from the easiness/difficulty in processing item in trial $n - 1$ cascading on the processing of the item in trial n could be implemented also in a dual-route architecture. To be able to reproduce the dynamic changes induced by list context on stimulus processing, a dual-route model ought to incorporate a thresholded mode of processing (as suggested by O'Malley and Besner, 2008) along with a separate control mechanism modulating route-change procedures (see Reynolds and Besner, 2008) Not even the dynamic approach provided by the ASE model is currently able to account for all the effects found in behavioral data. However, it offers a promising perspective for capturing the peculiarity of human cognition, i.e., flexibility and strategic behavior.

ACKNOWLEDGMENTS

This publication was supported by a grant from the Catholic University of Milan in its program for the promotion and dissemination of scientific research (D.3.1 2014).

REFERENCES

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

Baluch, B., and Besner, D. (1991). Strategic use of lexical and nonlexical routines in visual word recognition: evidence from oral reading in Persian. *J. Exp. Psychol. Learn. Mem. Cogn.* 17, 252–259. doi: 10.1037/0278-7393.17.4.644

Bodner, G. E., and Masson, M. E. J. (2003). Beyond spreading activation: an influence of relatedness proportion on masked semantic priming. *Psychon. B Rev.* 10, 645–652. doi: 10.3758/BF03196527

Borowsky, R., and Besner, D. (1993). Visual word recognition: a multistage activation model. *J. Exp. Psychol. Learn. Mem. Cogn.* 19, 813–840. doi: 10.1037/0278-7393.19.4.813

Chateau, D., and Lupker, S. J. (2003). Strategic effects in word naming: examining the route-emphasis versus time-criterion accounts. *J. Exp. Psychol. Hum. Percept. Perform.* 29, 139–151. doi: 10.1037/0096-1523.29.1.139

Coltheart, M. (1978). "Lexical access in simple reading tasks," in *Strategies of Information Processing*, ed. G. Underwood (New York: Academic Press), 131–216.

Coltheart, M., Davelaar, E., Jonasson, J. F., and Besner, D. (1977). "Access to the internal lexicon," in *Attention and Performance VI*, ed. S. Dornic (Hillsdale, NJ: Erlbaum), 535–555.

Coltheart, M., and Rastle, K. (1994). Serial processing in reading aloud: evidence for dual-route models of reading. *J. Exp. Psychol. Hum. Percept. Perform.* 20, 1197–1211. doi: 10.1037/0096-1523.20.6.1197

Coltheart, M., Rastle, K., Perry, C., Langdon, R., and Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychol. Rev.* 108, 204–256. doi: 10.1037//0033-295X.108.1.204

Goswami, U., and Ziegler, J. C. (2006). Fluency, phonology and morphology: a response to the commentaries on becoming literate in different languages. *Dev. Sci.* 9, 451–453. doi: 10.1111/j.1467-7687.2006.00511.x

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

Kinoshita, S., and Lupker, S. J. (2002). Effects of filler type in naming: change in time criterion or attentional control of pathways? *Mem. Cognit.* 30, 1277–1287. doi: 10.3758/BF03213409

Kinoshita, S., and Lupker, S. J. (2003). Priming and attentional control of lexical and sublexical pathways in naming: a reevaluation. *J. Exp. Psychol. Learn. Mem. Cognit.* 29, 405–415. doi: 10.1037/0278-7393.29.3.405

Kinoshita, S., Mozer, M. C., and Forster, K. I. (2011). Dynamic adaptation to history of trial difficulty explains the effect of congruency proportion on masked priming. *J. Exp. Psychol. Gen.* 140, 622–636. doi: 10.1037/a0024230

Lupker, S. J., Brown, P., and Colombo, L. (1997). Strategic control in a naming task: changing routes or changing deadlines? *J. Exp. Psychol. Learn. Mem. Cogn.* 23, 570–590. doi: 10.1037/0278-7393.23.3.570

Masson, M. E. J., and Kliegl, R. (2013). Modulation of additive and interactive effects in lexical decision by trial history. *J. Exp. Psychol. Learn. Mem. Cogn.* 39, 898–914. doi: 10.1037/a0029180

Monsell, S., Patterson, K. E., Graham, A., Hughes, C. H., and Milroy, R. (1992). Lexical and sublexical translation of spelling to sound: strategic anticipation of lexical status. *J. Exp. Psychol. Learn. Mem. Cogn.* 18, 452–467. doi: 10.1037/0278-7393.18.3.452

O'Malley, S., and Besner, D. (2008). Reading aloud: qualitative differences in the relation between stimulus quality and word frequency as a function of context. *J. Exp. Psychol. Learn. Mem. Cogn.* 34, 1400–1411. doi: 10.1037/a0013084

O'Malley, S., and Besner, D. (2013). Reading aloud: does previous trial history modulate the joint effects of stimulus quality and word frequency? *J. Exp. Psychol. Learn. Mem. Cogn.* 39, 1321–1325. doi: 10.1037/a0031673

O'Malley, S., Reynolds, M. G., and Besner, D. (2007). Qualitative differences between the joint effects of stimulus quality and word frequency in reading aloud and lexical decision: extensions to Yap and Balota. *J. Exp. Psychol. Learn. Mem. Cogn.* 33, 451–458. doi: 10.1037/0278-7393.33.2.451

Pagliuca, G., Arduino, L. S., Barca, L., and Burani, C. (2008). Fully transparent orthography, yet lexical reading aloud: the lexicality effect in Italian. *Lang. Cogn. Proc.* 23, 422–433. doi: 10.1080/01690960701626036

Paizi, D., Burani, C., De Luca, M., and Zoccolotti, P. (2011). List context manipulation reveals orthographic deficits in Italian readers with developmental dyslexia. *Child Neuropsychol.* 17, 459–482. doi: 10.1080/09297049.2010.551187

Paizi, D., Burani, C., and Zoccolotti, P. (2010). List context effects in reading Italian words and nonwords: can the word frequency effect be eliminated? *Eur. J. Cogn. Psychol.* 22, 1039–1065. doi: 10.1080/09541440903216492

Paulesu, E., McCrory, E., Fazio, F., Menoncello, L., Brunswick, N., Cappa, F., et al. (2000). A cultural effect on brain function. *Nat. Neurosci.* 3, 91–96. doi: 10.1038/71163

Perry, C., Ziegler, J. C., and Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: the CDP+ model of reading aloud. *Psychol. Rev.* 114, 273–315. doi: 10.1037/0033-295X.114.2.273

Perry, C., Ziegler, J. C., and Zorzi, M. (2010). Beyond single syllables: large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. *Cogn. Psychol.* 61, 106–151. doi: 10.1016/j.cogpsych.2010.04.001

Perry, C., Ziegler, J. C., and Zorzi, M. (2014). When silent letters say more than a thousand words: an implementation and evaluation of CDP++ in French. *J. Mem. Lang.* 72, 98–115. doi: 10.1016/j.jml.2014.01.003

Plaut, D. C., and Booth, J. R. (2000). Individual and developmental differences in semantic priming: empirical and computational support for a single-mechanism account of lexical processing. *Psychol. Rev.* 107, 786–823. doi: 10.10037/0033-295X.107.4.786

Plaut, D. C., and Booth, J. R. (2006). More modeling but still no stages: reply to Borowsky and Besner. *Psychol. Rev.* 113, 196–200. doi: 10.1037/0033-295X.113.1.196

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., and Patterson, K. E. (1996). Understanding normal and impaired reading: computational principles in quasi-regular domains. *Psychol. Rev.* 103, 56–115. doi: 10.1037/0033-295X.103.1.56

Pritchard, S. C., Coltheart, M., Palethorpe, S., and Castles, A. (2012). Nonword reading: comparing dual-route cascaded and connectionist dual-process models with human data. *J. Exp. Psychol. Hum. Percept. Perform.* 38, 1268–1288. doi: 10.1037/a0026703

Raman, I., Baluch, B., and Besner, D. (2004). On the control of visual word recognition: changing routes versus changing deadlines. *Mem. Cognit.* 32, 489–500. doi: 10.3758/BF03195841

Rastle, K., and Coltheart, M. (1999). Serial and strategic effects in reading aloud. *J. Exp. Psychol. Hum. Percept. Perform.* 25, 482–503. doi: 10.1037/0096-1523.25.2.482

Reynolds, M., and Besner, D. (2008). Contextual effects on reading aloud: evidence for pathway control. *J. Exp. Psychol. Learn. Mem. Cognit.* 34, 50–64. doi: 10.1037/0278-7393.34.1.50

Scaltritti, M., Balota, D. A., and Peressotti, F. (2013). Exploring the additive effects of stimulus quality and word frequency: the influence of local and list-wide prime relatedness. *Q. J. Exp. Psychol.* 66, 91–107. doi: 10.1080/17470218.2012.698628

Seidenberg, M. S., and McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychol. Rev.* 96, 523–568. doi: 10.1037/0033-295X.96.4.523

Seidenberg, M. S., Plaut, D. C., Petersen, A. S., McClelland, J., and McCrae, K. (1994). Nonword pronunciation and models of word recognition. *J. Exp. Psychol. Hum. Percept. Perform.* 20, 1177–1196. doi: 10.1037/0096-1523.20.6.1177

Stanners, R. F., Jastrzembski, J. E., and Westbrook, A. (1975). Frequency and visual quality in a word-nonword classification task. *J. Verbal Learn.* 14, 259–264. doi: 10.1016/S0022-5371(75)80069-7

Sternberg, S. (1969). The discovery of processing stages: extensions of Donder's method. *Acta Psychol.* 30, 276–315. doi: 10.1016/0001-6918(69)90055-9

Tabossi, P., and Laghi, L. (1992). Semantic priming in the pronunciation of words in two writing systems: Italian and English. *Mem. Cognit.* 20, 303–313. doi: 10.3758/BF03199667

Underwood, G. (1978). "Concepts in information processing theory," in *Strategies of Information Processing*, ed. G. Underwood (New York: Academic Press), 1–22.

Yap, M. J., and Balota, D. A. (2007). Additive and interactive effects on response time distributions in visual word recognition. *J. Exp. Psychol. Learn. Mem. Cognit.* 33, 274–296. doi: 10.1037/0278-7393.33.2.274

Yates, M. (2010). Investigating the importance of the least supported phoneme on visual word naming. *Cognition* 115, 197–201. doi: 10.1016/j.cognition.2009.12.002

Zevin, J. D., and Balota, D. A. (2000). Priming and attentional control of lexical and sublexical pathways during naming. *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 121–135. doi: 10.1037//0278-7393.26.1.121

Ziegler, J. C., and Goswami, U. C. (2005). Reading acquisition, developmental dyslexia and skilled reading across languages: a psycholinguistic grain size theory. *Psychol. Bull.* 131, 3–29. doi: 10.1037/0033-2909.131.1.3

Ziegler, J. C., and Goswami, U. C. (2006). Becoming literate in different languages: similar problems, different solutions. *Dev. Sci.* 9, 429–436. doi: 10.1111/j.1467-7687.2006.00509.x

Ziegler, J. C., Perry, C., and Zorzi, M. (2009). Additive and interactive effects of stimulus degradation: no challenge for CDP+. *J. Exp. Psychol. Learn. Mem. Cogn.* 35, 306–311. doi: 10.1037/a0013738

Zorzi, M., Houghton, G., and Butterworth, B. (1998). Two routes or one in reading aloud? A connectionist dual-process model. *J. Exp. Psychol. Hum. Perform. Percept.* 24, 1131–1161. doi: 10.1037/0096-1523.24.4.1131

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.

Received: 14 March 2014; accepted: 27 August 2014; published online: 15 September 2014.

Citation: Traficante D and Burani C (2014) List context effects in languages with opaque and transparent orthographies: a challenge for models of reading. Front. Psychol. ***5****:1023. doi: 10.3389/fpsyg.2014.01023*

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

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 14 April 2014
doi: 10.3389/fpsyg.2014.00315

An ERP study of effects of regularity and consistency in delayed naming and lexicality judgment in a logographic writing system

Yen Na Yum[1], Sam-Po Law[1]*, I-Fan Su[1], Kai-Yan Dustin Lau[2] and Kwan Nok Mo[1]

[1] *Division of Speech and Hearing Sciences, University of Hong Kong, Hong Kong, China*
[2] *Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, China*

Edited by:
Davide Crepaldi, University of Milano-Bicocca, Italy

Reviewed by:
Marcus Taft, University of New South Wales, Australia
Stéphanie Massol, Basque Center on Cognition, Brain and Language, Spain

***Correspondence:**
Sam-Po Law, Division of Speech and Hearing Sciences, University of Hong Kong, Rm804B Meng Wah Complex, Pokfulam Road, Hong Kong, China
e-mail: splaw@hku.hk

Phonological access is an important component in theories and models of word reading. However, phonological regularity and consistency effects are not clearly separable in alphabetic writing systems. We investigated these effects in Chinese, where the two variables are operationally distinct. In this orthographic system, regularity is defined as the congruence between the pronunciation of a complex character (or phonogram), and that of its phonetic radical, while phonological consistency indexes the proportion of orthographic neighbors that share the same pronunciation as the phonogram. In the current investigation, regularity and consistency were contrasted in an event-related potential (ERP) study using a lexical decision (LD) task and a delayed naming (DN) task with native Chinese readers. ERP results showed that effects of regularity occurred early after stimulus onset and were long-lasting. Regular characters elicited larger N170, smaller P200, and larger N400 compared to irregular characters. In contrast, significant effects of consistency were only seen at the P200 and consistent characters showed a greater P200 than inconsistent characters. Thus, both the time course and the direction of the effects indicated that regularity and consistency operated under different mechanisms and were distinct constructs. Additionally, both of these phonological effects were only found in the DN task and absent in LD, suggesting that phonological access was non-obligatory for LD. The study demonstrated cross-language variability in how phonological information was accessed from print and how task demands could influence this process.

Keywords: phonological regularity, phonological consistency, Chinese, delayed naming, lexical decision, event-related potential (ERP)

INTRODUCTION

All writing systems carry phonological information, but they vary in the nature of correspondence between orthographic units, e.g., whole word, sublexical components, and phonological units, e.g., phonemes, rimes, syllables. For instance, in alphabetic scripts such as English, French, German, and Korean *hangul*, the orthography-phonology mapping is between letters and phonemes; the correspondence in systems such as Japanese *katakana* and *hiragana* is between a symbol (i.e., kana) and a syllable (or more precisely mora); and in the case of Chinese, each character, considered a logograph by some, is associated with a syllable. These cross-linguistic variations are expected to have profound impact on how phonological information is accessed from print and implications for models of reading.

Given the existence of sublexical correspondence, all theoretical models of reading in alphabetic scripts assume a nonlexical reading mechanism without necessarily a lexical route (e.g., Coltheart, 1978; Hillis and Caramazza, 1995; Plaut et al., 1996). The orthography-phonology relationships can be characterized in terms of regularity and consistency, depending on the theoretical approach. The regularity of a word is determined by whether its pronunciation conforms to grapheme-phoneme correspondence (GPC) rules of the language (e.g., regular words such as *raid*, *pink* vs. irregular words such as *pint*, *have*; Coltheart et al., 1993, 2001), while the consistency of a word depends on the strength of spelling-sound connections derived from the properties of the pronunciations of the "body" of other similarly spelled words (e.g., consistent words such as *bust*, *dust*, *gust*, *just*, *lust*, *must*, *rust* vs. inconsistent words such as *cost*, *host*, *lost*, *most*, *post*; Seidenberg and McClelland, 1989; Plaut et al., 1996). Both regularity and consistency have been shown to affect naming latency. Irregular words take longer to name than regular words (e.g., Baron and Strawson, 1976; Gough and Cosky, 1977; Stanovich and Bauer, 1978), and the effect is more pronounced in low frequency words (e.g., Andrews, 1982; Seidenberg et al., 1984; Waters et al., 1984). Readers are also slower to read aloud inconsistent than consistent lexical items (Glushko, 1979). However, regularity and consistency are not easily distinguishable. Irregular or exception words are often inconsistent; moreover, in some studies regularity is defined in terms of neighborhood characteristics such as the relative numbers of friends (e.g., peak-teak) and enemies (e.g., peak-pear) (e.g., Peereman, 1995). In the few studies that have manipulated both regularity and consistency, effects of consistency are robust while those of regularity are unclear or limited (Andrews, 1982; Kay and Bishop, 1987; Cortese and Simpson, 2000; Jared, 2002). This has raised the question whether

regularity effects conceptualized as GPC knowledge have important impact on reading alphabetic scripts. Interestingly, although the Chinese writing system is generally considered logographic, the notions of regularity and consistency have been shown to be highly relevant to reading, and they are theoretically more distinct by comparison. Hence, the present study investigated their underlying mechanism using a technique known for its excellent temporal resolution and ability to reveal online unfolding of cognitive processes, i.e., event-related potential (ERP), in addition to the traditional behavioral measures.

Almost all Chinese characters are monosyllabic and correspond to morphemes. As such, the Chinese script is described as a morphosyllabic system. Given that there are no elements within a character that correspond to phonemes or tone, the postulation of a non-lexical reading pathway in Chinese may be irrelevant. Nonetheless, more than 80% of all Chinese characters are phonograms consisting of components that carry some semantic and phonological information of the character. Orthographically, Chinese characters are made up of spatial arrangements of strokes, which combine to form larger units called "radicals." Radicals may further combine to form complex characters or phonograms. Phonograms contain a semantic radical and a phonetic radical providing a clue to the meaning of a character and one to the pronunciation of the character, respectively. For instance, the character 趾 *zi2* "toe" has a semantic radical 足 on the left meaning "foot" and a phonetic radical 止 *zi2* on the right. (In this paper, phonetic transcriptions of Chinese characters are given in *jyutping*, a romanization system developed by the Linguistics Society of Hong Kong. The number in the transcription represents the tone.) According to the entries in two dictionaries of Cantonese phonograms (Ni, 1982; Li, 1989), Law et al. (2009) reported that about 34–40% of phonograms are "regular" characters. Their pronunciations are segmentally identical (regardless of tone) to the pronunciation of their phonetic radical when it occurs as a character (e.g., 湖 *wu4* and 胡 *wu4*). Another 30% are "partially regular" phonograms sharing at least the same rime as their phonetic radical (e.g., 他 *taa1* and 也 *jaa5*), and the rest are "irregular" with no phonological relationship with their phonetic radical (e.g., 路 *lou6* and 各 *gok3*). Most phonetic radicals are also existing characters, such as 止, 胡, 也, 各; however, it is important to note that there is a non-negligible number of phonetic radicals, approximately 16% of all phonetic radical entries listed in Li (1989), that do not exist alone, e.g., the right-hand components of 疏, 搖, 蛙. These radicals have no associated phonological representations or meaning.

Besides regularity, the phonological property of a character can also be described in terms of consistency. It refers to the extent to which the phonetic radical serves as a reliable cue to the pronunciations of the phonograms containing it. A character of high consistency is one that sounds the same as most, if not all phonograms sharing the same phonetic radical (e.g., 驅 *keoi1*, 軀 *keoi1*, 嶇 *keoi1*, 毆 *keoi1*, 鱋 *keoi1*, 嘔 *ngau2*), and a low consistency character is one that shares the same phonetic radical with phonograms that sound differently (e.g., 油 *jau4*, 宙, *zau6*, 迪 *dik6*, 笛 *dek6*, 軸 *zuk6*). In other words, regularity is defined by the phonological distance between a phonogram and its phonetic radical and only applicable to phonograms with phonetic radicals that exist as stand alone characters, while consistency is determined by the number of different phonological forms associated with a family or neighborhood of phonograms having a common phonetic radical. One can see that consistency in Chinese is comparable to that in alphabetic scripts (Lee, 2008), whereas regularity has a distinct definition.

Psycholinguistic studies of character recognition has accumulated ample evidence that phonetic radicals access phonological representations independently of and in parallel with the phonograms (e.g., Seidenberg, 1985; Hue, 1992; Wu et al., 1994; Weekes et al., 1998; Zhou and Marslen-Wilson, 1999; Ding et al., 2004). Low frequency regular phonograms, but not high frequency ones, have significantly shorter reading latencies than irregular phonograms. Similarly, in a series of reading aloud experiments, Fang et al. (1986) demonstrated effects of consistency. Regular/consistent characters were named significantly faster than regular/inconsistent and irregular phonograms. Comparable findings were reported in Lian (1985) and more recently in Lee et al. (2005). Lee et al. manipulated character frequency, regularity and consistency. A significant interaction between regularity and consistency was found for low frequency characters; furthermore, consistency effects were observed among irregular but not regular phonograms. Irregular high consistency characters were named faster than irregular low consistency characters. These findings suggest that regularity and consistency are independent variables. It is, however, worth noting that these observations were based on a very small set of stimuli of 10 in each experimental condition.

Besides behavioral evidence, the "ortho-phonological" effects can also be observed in neural responses of specific ERP components. Using the homophone judgment task in which stimuli contained phonetic radicals varying in consistency, Lee and colleagues found effects of consistency in N170 in the temporo-occipital region, P200 in the frontal region, and N400 in the central region. In particular, greater negativity in N170 and greater positivity in P200 were elicited by inconsistent characters, compared with consistent ones, while greater N400 was found for consistent than inconsistent characters (Lee et al., 2007). The effects at N170 and P200 were interpreted as early extraction of phonological information from the phonetic radical, whereas effects at N400 were taken to reveal post-lexical processing resulting from competition among activated representations at the lexical level. Somewhat different results were reported when pseudo-characters were employed. Although pseudo-characters containing unpredictable (inconsistent) phonetic radicals exhibited greater P200, they also elicited greater N400 (Lee et al., 2006a). When consistency was manipulated with neighborhood characteristics taken into consideration, including orthographic neighborhood (number of phonograms sharing the same phonetic radical), phonological alternatives (number of different phonological forms associated with a phonetic radical family), and phonological neighborhood (number of homophonic characters associated with a phonological form), a more fine-grained picture emerged (Hsu et al., 2009). High consistency characters exhibited greater negativity in N170, smaller P200, and greater N400 compared with low consistency stimuli, but these effects

were restricted to characters from large orthographic neighborhoods ($N > 10$) compared with small ones ($N < 4$).

It is notable that the aforementioned studies focused on the consistency effect. The only ERP study that has involved regularity was Lee et al. (2006b). In a character recognition task, participants were presented with pairs of prime-target characters varying in stimulus-onset-asynchrony (SOA) and semantic relatedness. Among the three conditions in which the prime and target are semantically unrelated, two of them involved a phonogram prime containing a phonetic radical that is semantically related to the target but these two conditions differed in terms of whether the prime was a regular or irregular phonogram (e.g., Regular phonogram: 楓 (prime) *fung1* "maple" (phonetic radical 風 *fung1* "wind") → 雨 (target) *jyu5* "rain"; Irregular phonogram: 讀 *duk6* "read" (賣 *maai6* 'sell') → 買 *maai5* "buy"). Both conditions revealed significant N400 semantic priming effects when contrasted with the unrelated control condition but only in the shorter SOA conditions (50 and 100 ms). Moreover, the N400 effect elicited by the regular phonograms appeared earlier and persisted longer than the irregular phonograms. These results suggest that the phonological forms of the phonogram and its phonetic radical have modulating effects on semantic processing during the N400 time window.

In summary, few ERP studies have examined regularity and consistency simultaneously and how they may differ in neural representation. Moreover, the contrast in consistency in previous work was often between extreme values, especially for high consistency characters with an average consistency approaching 1. Little information was provided on the composition of the high and low consistency characters with respect to their regularity status. In other words, it is not clear whether stimuli in the two consistency conditions had comparable number of regular and irregular characters. Hence, consistency might have been confounded with regularity in previous ERP works.

Given the conceptual distinction between "word-based" regularity and "neighborhood-based" consistency, it is reasonable to expect that they differ in neural correlates, at least in terms of time course. To illustrate, upon seeing a medium-to-high frequency phonogram containing a free-standing phonetic radical, a skilled reader may be able to immediately segment the character into radical components and access the corresponding phonological forms, i.e., the whole character and the phonetic radical. The phonetic radical then spreads activation to phonograms containing it; the activated phonograms then access their phonological representations, which compete with one another. Such a scenario is compatible with most models of reading in Chinese (e.g., Taft and Zhu, 1997; Perfetti et al., 2005). It also predicts that the regularity effect may emerge earlier *and* last longer than the consistency effect. The former effect results from competition between two phonological forms activated by direct orthography-to-phonology mapping, while the latter arises from competition among phonograms activated by the segmented phonetic radical. Moreover, we hypothesize that the consistency effect has a shorter time course than regularity. Competition between a phonogram and its phonetic radical is driven by orthographic forms of the stimulus, and therefore, persists until a selection for output is made. In contrast, phonograms in an orthographic neighborhood are activated "indirectly" by the phonetic radical in the stimulus, and the majority of the activated representations do not correspond to the target. These predictions differ importantly from previous findings by Lee and colleagues, which would predict consistency effects in the time windows of N170, P200, and N400, and regularity effects occurring mainly in N400.

The current study employed behavioral and neural measures of regularity and consistency. Given the impact of the characteristics of orthographic and phonological neighborhoods on character naming, and the difficulties in identifying enough lexical items varying in regularity and consistency while matched on neighborhood variables, effects of regularity and consistency were studied separately using different sets of stimuli. In addition to using a task that explicitly accesses phonological information, i.e., reading aloud characters but after a delay to eliminate movement artifacts undesirable in ERP experiments, a lexical decision (LD) task was administered. Lexicality judgment is probably the most common task in lexical processing research. While not central to our research questions, performance in lexicality judgment ensures that participants attend to the stimuli and the experimental task. Previous studies have shown enhanced N400 to pseudowords compared to real words, interpreted as reflecting difficulty in lexical access (Bentin et al., 1985; Holcomb, 1993; Nobre and McCarthy, 1994). Although lexicality can be determined without recourse to phonology, the presence or absence of phonological effects in such as task has both theoretical and practical significance. Most theoretical models assume that access to phonology is automatic upon seeing a written word without reference to the goal(s) of a task. A comparison between reading aloud and LD will allow us to see if the reading processes involved change as a function of task demands. If phonological information is available automatically in lexicality judgment and the effects are of comparable strength to naming, then LD would be preferred especially in reading experiments using ERPs, because responses are based on single stimuli as opposed to pairs of stimuli in homophone judgment and relatively free of motion artifacts.

MATERIALS AND METHODS

PARTICIPANTS

Twenty four (12 females) right-handed native Cantonese speakers aged 18–26 ($M = 21.17$, $SD = 1.97$) with normal neurological profile and visual acuity were recruited for this study. Participants received cash compensation upon completion of experimental tasks. Written informed consent was obtained from all participants and the experiments were approved by the Human Research Ethics Committee for Non-Clinical Faculties of the University of Hong Kong.

MATERIALS

Real word stimuli consisted of 160 phonograms written in traditional Chinese script of left-right or top-bottom configuration with one phonetic and one semantic radical. In the LD task, 160 pseudo-characters were used as well. These were created by randomly combining the phonetic and semantic radicals of the real character stimuli in accordance to orthographic rules.

Pseudo-characters and real characters were matched on structural configuration and stroke number.

Two lists of characters selected from the real words were used to investigate the effects of phonological regularity and consistency in LD and Delayed Naming (DN) tasks. Regularity was defined by the relationship between the pronunciation of the character and that of its phonetic radical. Regular characters shared both onset and rime with its phonetic radical regardless of tone, while irregular characters did not meet this criterion. One common method to calculate a consistency value is to divide the number of friends (orthographic neighbors that share the same pronunciation) by the total number of orthographic neighbors. This proportion is known as type consistency. Another way of measuring consistency, known as token consistency, takes into account the lexical frequency of the orthographic neighbors, i.e., giving more weight to neighbors with higher frequency. Note that the measure of regularity is only meaningful when the phonetic radical is an existing character that carries its own pronunciation, while consistency does not have this limitation. In addition, the phonetic radical was not counted as a neighbor in our calculations of type or token consistency values; this is different from the estimates of consistency in Lee and colleagues' work.

In our stimuli, regular and irregular characters ($n = 55$ in each condition) were matched in stroke number, lexical frequency, orthographic neighborhood size, type and token phonological consistency, number of homophones, number of syllables associated with an orthographic neighborhood, and lexical frequency of the phonetic radical. Consistent and inconsistent characters ($n = 36$ in each condition) were significantly different in type as well as token consistency. They also differed in the number of phonological alternatives for the phonetic radical, as inconsistent characters tended to have more orthographic neighbors with different pronunciations. Importantly, for both consistent and inconsistent items, half were regular and half were irregular. They were also matched in stroke number, lexical frequency, orthographic neighborhood size, number of homophones, and standalone frequency of the phonetic radical. Properties of the stimuli used in each condition are shown in **Table 1**.

PROCEDURE

After informed consent, participants were seated in an electrically and acoustically shielded room. Stimuli were presented on a computer screen located approximately 60 cm away. For all participants, a LD was administered followed by a delayed naming task (DN). A practice block was given to each participant prior to each task. On each trial of LD, a fixation cross (500 ms) preceded a yellow character (100 × 90 pixels) was presented for 800 ms (1200–1500 ms ITI) on a black background. Participants decided if the character was real by pressing a button for real characters and another button for pseudo-characters. The stimuli were given in six blocks in a random sequence delivered by E-Prime (Psychology Software Tools Inc., USA), with the response buttons counterbalanced across participants.

In DN, only real characters were shown. On each trial, a fixation cross was presented for 500 ms and the character was displayed for 800 ms. Then the character would be replaced by three asterisks, which remained on the screen until a response was made. Participants were instructed to name the displayed character upon seeing the asterisks. ERP measurement was time-locked to the visual onset of characters, prior to actual utterance. The response delay served to reduce muscle artifacts produced during verbal production. The responses were recorded and coded offline for response accuracy.

EEG recordings

The EEG data were recorded from 64 Ag/AgCl electrodes (10–20 system) with a common vertex reference electrode located between electrodes Cz and CPz, and ground (GND) positioned anterior to electrode Fz. Vertical and horizontal eye movements were monitored by bipolar electrodes (VEOG) placed on the supra- and infraorbital ridges of the left eye and bipolar electrodes (HEOG) placed on the left and right side of the lateral orbital rim. Electrode impedance was maintained below 5 KΩ and data were digitized online at 1 kHz with a band pass filter of 0.05–200 Hz using SynAmps2® (Neuroscan, Inc., El Paso, TX, USA) amplifiers.

ERP data processing

In the off-line analysis, continuous data were filtered using a zero phase shift low-pass filter of 30 Hz (12 dB/octave slopes). Channels affected by eye blink artifacts were corrected using a model artifact implemented in Scan 4.5 software (Neuroscan, Inc), with a minimum of 100 eyeblink artifacts for each participant. Segments of −200 to 1000 ms post-stimulus onset intervals were later extracted and baseline corrected using the pre-stimulus intervals (−200 to 0 ms). Trials with incorrect responses, muscle artifacts, or voltage exceeding 100 μV were automatically rejected. The remaining data were re-referenced to the average of the two mastoid electrodes and used to compute grand average waveforms for each condition.

STATISTICAL ANALYSES

For behavioral effects of lexicality, *t*-tests were used to compare accuracy and response time (RT) to real and pseudo-characters in LD. For consistency and regularity effects, *t*-tests were performed on the accuracy and RT data in LD. Since a response delay was introduced in DN, only effects on naming accuracy were examined in this task.

In both LD and DN, mean amplitudes of the N170, P200, and N400 ERP components timelocked to character onset were examined and analyzed statistically. Three-Way ANOVAs were conducted for each of the three components, with Electrode Location (N170: P5, P6, P7, P8, PO5, PO6, PO7, PO8; P200: FC3, FC4, C3, C4, CP3, CP4; N400: FC5, FC6,C5, C6, CP5, CP6,P5, P6) and Hemisphere (left vs. right) as within-subject independent variables in addition to Tasks (LD vs. DN) and Experimental Conditions (Consistent vs. Inconsistent or Regular vs. Irregular). We chose these electrode locations based on previous ERP works on these phonological effects in Chinese (Lee et al., 2007; Hsu et al., 2009). Estimation of analysis windows was based on the peak latencies derived from the mean amplitude for all trials at the selected electrode locations. The window for N170 was set as 100–200 ms, with the peak at 151 ms. The P200 window was 200–270 ms, with the peak at 233 ms. The N400 window was 270–400 ms, with the peak at 326 ms. These

Table 1 | Properties of the stimuli in the regularity and the consistency contrasts.

	Irregular ($N = 55$)		Regular ($N = 55$)		
	Range	**Mean (*SD*)**	**Range**	**Mean (*SD*)**	***p*-value**
Stroke	6–20	11.76 (2.96)	6–20	11.58 (3.34)	0.76
Frequency (per mil.)	0.31–1306.43	333.67 (386.29)	4.23–2075.05	325.87 (419.34)	0.92
Family size	3–15	7.16 (3.18)	3–15	6.96 (3.05)	0.74
Consistency (Type)	0.07–0.75	0.33 (0.20)	0.07–1	0.35 (0.26)	0.61
Consistency (Token)	0.01–0.99	0.5 (0.36)	0.01–1	0.45 (0.36)	0.48
No. of homophones	0–16	4.42 (3.98)	0–22	4.67 (4.82)	0.76
No. of associated syllables	2–12	4.40 (2.20)	1–8	4.40 (1.99)	1
Radical frequency (per mil.)	3.13–5616.44	607.58 (1121.61)	11.98–4090.13	788.91 (944.54)	0.36
	Consistent ($N = 36$)		**Inconsistent ($N = 36$)**		
Stroke	7–20	12.44 (3.26)	7–20	12.08 (3.76)	0.66
Frequency (per mil.)	5.92–1299.44	237.79 (331.07)	7.93–1053.11	306.56 (331.53)	0.38
Family size	5–8	6.31 (0.86)	4–9	6.56 (1.38)	0.36
Consistency (Type)	0.13–1	0.53 (0.28)	0.13–0.50	0.21 (0.09)	<0.001
Consistency (Token)	0.09–1	0.71 (0.32)	0.01–0.80	0.22 (0.21)	<0.001
No. of homophones	0–15	4.81 (3.62)	0–19	4.06 (4.93)	0.46
No. of associated syllables	1–4	2.75 (1.05)	4–8	5.22 (0.93)	<0.001
Radical frequency (per mil.)	0.63–4090.13	665.70 (1039.60)	0–1917	408.15 (495.87)	0.19

time windows were roughly comparable to previous findings (Lee et al., 2007; Hsu et al., 2009). The lexicality effect in LD was examined with all real characters and pseudo-characters at the N400 component using the same electrode locations and time window. The significance threshold for *post-hoc* ANOVAs was corrected for multiple comparisons using Bonferroni adjustment.

RESULTS

BEHAVIORAL RESULTS

A summary of the behavioral findings is shown in **Table 2**. In LD, trials with response latencies below 200 ms and exceeding 2000 ms were discarded (<1%), and incorrect trials were excluded in RT calculations. For lexicality effects in LD, participants responded more accurately to real characters ($M = 97\%$, $SD = 0.02$) than to pseudo-characters [$M = 89\%$, $SD = 0.09$; $t_{1(23)} = 3.53$, $p = 0.002$; $t_{2(159)} = 7.12$, $p < 0.001$ where t_1 denotes results from subject analyses and t_2 denotes results from item analyses]. Participants were also faster when responding to real characters ($M = 546$ ms, $SD = 58.50$) compared to pseudo-characters [$M = 661$ ms, $SD = 123.98$; $t_{1(23)} = 6.07$, $p < 0.001$; $t_{2(159)} = 22.94$, $p < 0.001$].

For effects of regularity and consistency on response latencies in LD, a marginal effect of faster RT to regular characters than irregular characters was found in the subject analysis, but this was not significant in the item analysis [$t_{1(23)} = 1.74$, $p = 0.096$; $t_{2(54)} = 0.78$, $p = 0.441$]. Regularity did not have a significant effect on lexicality judgment accuracy [$t_{1(23)} = 0.58$, $p = 0.567$; $t_{2(54)} = 0.37$, $p = 0.714$]. Participants were marginally slower when responding to consistent characters than to inconsistent characters, again only in the subject analysis [$t_{1(23)} = 1.82$, $p = 0.081$; $t_{2(35)} = 0.90$, $p = 0.373$]. Response accuracy was higher for inconsistent than consistent characters, but this was only significant in the subject analysis [$t_{1(23)} = 2.33$, $p = 0.029$; $t_{2(35)} = 1.52$, $p = 0.136$].

In DN, higher accuracy for regular characters than irregular characters was revealed in the subject analysis, but not the item analysis [$t_{1(23)} = 2.22$, $p = 0.037$; $t_{2(54)} = 0.828$, $p = 0.411$]. Higher naming accuracy for consistent than inconsistent characters was shown in the subject analysis only [$t_{1(23)} = 2.94$, $p = 0.007$; $t_{2(35)} = 1.12$, $p = 0.270$].

In short, effects of lexicality on both response accuracy and latency were significant. In contrast, none of the results of the regularity and consistency contrasts were statistically reliable.

ERP RESULTS

On average, 10.4% of trials were rejected due to incorrect responses or other artifacts. LD had more remaining trials than DN in both the regularity contrast ($M = 50.5$ vs. 47.8) and the consistency contrast ($M = 33.5$ vs. 31.1). However, the numbers of trials for regular and irregular characters and for consistent and inconsistent characters were comparable in each task. The grand average waveforms and voltage maps for the consistency and regularity contrasts at N170, P200, and N400 time windows are plotted in **Figures 1–6**. Those for the lexicality contrast at the N400 window are shown in **Figures 7**, **8**, respectively.

N170 (100–200 ms)

A significant regularity × task interaction was seen in this component [$F_{(1,\ 23)} = 9.24$, $p = 0.006$, $\eta\rho^2 = 0.29$). A larger N170 was seen in regular characters than irregular characters in DN only (regular: $M = -0.68$, $SE = 0.53$; irregular: $M = -0.32$, $SE = 0.60$, $p = 0.014$) but not in LD (regular: $M = -0.47$, $SE = 0.43$; irregular: $M = -0.51$, $SE = 0.51$, $p > 0.05$), see **Figure 1A** and

Table 2 | Behavioral results in lexical decision and delayed naming, standard deviations are given in parentheses.

	Regular	Irregular	Consistent	Inconsistent
RT in LD (ms)	546 (55)	553 (66)	551 (65)	542 (54)
Accuracy in LD (%)	97.3 (3.1)	97.0 (2.8)	96.2 (3.0)	98.1 (2.8)
Accuracy in DN (%)	98.2 (2.1)	97.3 (2.1)	99.3 (1.2)	97.9 (2.2)

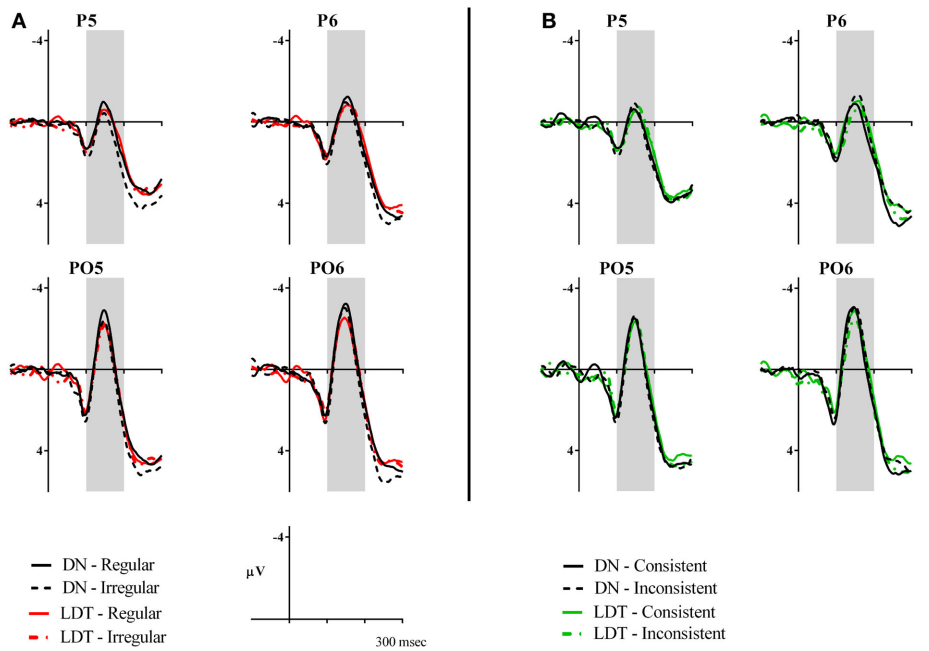


FIGURE 1 | (A) Grand average waveforms showing the N170 regularity effect at parietal and parietal-occipital electrodes (P5, P6, PO5, and PO6). Even numbers refer to electrode positions on the right hemisphere, whereas odd numbers refer to those on the left hemisphere. **(B)** Grand average waveforms showing the N170 consistency effect at the same electrodes. Shaded areas represent the analysis window of 100–200 ms.

Figure 2. Consistency effects were also observed in a consistency x task x hemisphere interaction [$F_{(1,\ 23)} = 6.90$, $p = 0.015$, $\eta\rho^2 = 0.23$]. Examination of **Figures 1B**, **2** suggested consistent characters elicited a larger N170 than inconsistent characters in the right hemisphere in LD. However, follow-up *post-hoc* analyses did not reveal significant two-way interactions or pairwise differences in any of the conditions.

P200 (200–270 ms)

A main effect of regularity was found [$F_{(1,\ 23)} = 4.64$, $p = 0.042$, $\eta\rho^2 = 0.17$], with irregular characters ($M = 2.90$, $SE = 0.32$) eliciting significantly larger P200 than regular characters ($M = 2.67$, $SE = 0.33$). This regularity effect was marginally modulated by task [$F_{(1,\ 23)} = 3.56$, $p = 0.072$, $\eta\rho^2 = 0.13$]. Irregular characters were more positive than regular characters in DN (regular: $M = 2.94$, $SE = 0.39$; irregular: $M = 3.50$, $SE = 0.40$, $p = 0.012$) but not in LD (regular: $M = 2.39$, $SE = 0.33$; irregular: $M = 2.30$, $SE = 0.33$, $p > 0.05$), see **Figures 3A, 4**. As for effects of consistency, a consistency x task interaction was seen in this time window [$F_{(1,\ 23)} = 4.97$, $p = 0.036$, $\eta\rho^2 = 0.18$]. Participants showed larger P200 in response to consistent characters than inconsistent characters in DN (consistent: $M = 3.56$, $SE = 0.42$; inconsistent: $M = 2.95$, $SE = 0.37$, $p = 0.024$) but not in LD (consistent: $M = 2.65$, $SE = 0.37$; inconsistent: $M = 2.62$, $SE = 0.35$, $p > 0.05$), see **Figures 3B**, **4**. Furthermore, a consistency × hemisphere × electrode interaction was found [$F_{(2,\ 46)} = 4.87$, $p = 0.016$, $\eta\rho^2 = 0.18$]. Although it appeared that the consistency effect was stronger in the right hemisphere, follow-up analyses did not reveal significant differences in two-way interactions or pairwise comparisons.

N400 (270–400 ms)

A main effect of regularity was obtained [$F_{(1,\ 23)} = 4.53$, $p = 0.044$, $\eta\rho^2 = 0.17$]. Significantly larger N400 was elicited by regular characters ($M = 2.12$, $SE = 0.26$) than irregular characters ($M = 2.32$, $SE = 0.30$), see **Figures 5A**, **6**. No other significant effects were obtained. There were also null effects of consistency in this component, see **Figures 5B**, **6**.

A main effect of lexicality was also seen in this window [$F_{(1,\ 23)} = 51.90$, $p < 0.001$, $\eta\rho^2 = 0.69$]. The N400 for pseudo-characters ($M = 1.57$, $SE = 0.29$) was much larger than for real characters ($M = 2.31$, $SE = 0.29$), see **Figures 7**, **8**. A

lexicality by hemisphere interaction was also observed [$F_{(1, 23)} = 4.84$, $p = 0.038$, $\eta\rho^2 = 0.17$]. *Post-hoc* comparisons showed that the lexicality effect was stronger at right hemisphere electrodes (real: $M = 2.81$, $SE = 0.34$; pseudo: $M = 1.81$, $SE = 0.36$, $p < 0.001$), but also significant at left hemisphere electrodes (real: $M = 1.80$, $SE = 0.34$; pseudo: $M = 1.33$, $SE = 0.33$, $p = 0.012$). Furthermore, lexicality interacted with electrode locations [$F_{(3, 69)} = 10.59$, $p < 0.001$, $\eta\rho^2 = 0.32$], but *post-hoc* analyses revealed significant differences at all electrode sites (all $p < 0.001$).

DISCUSSION

The current investigation examined the independence of regularity and consistency effects during Chinese character recognition

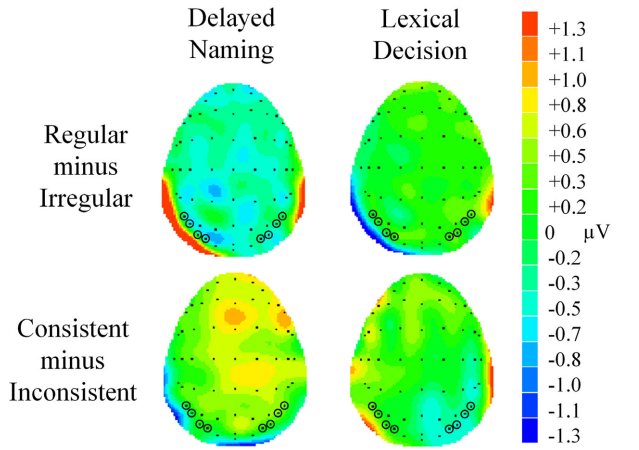


FIGURE 2 | Voltage maps showing effects of regularity and consistency in delayed naming and lexical decision in N170. The maps were calculated from the mean difference values (in μV) of regular minus irregular characters and consistent minus inconsistent characters between 100–200 ms. Electrodes P5, P6, P7, P8, PO5, PO6, PO7, and PO8 are circled for reference.

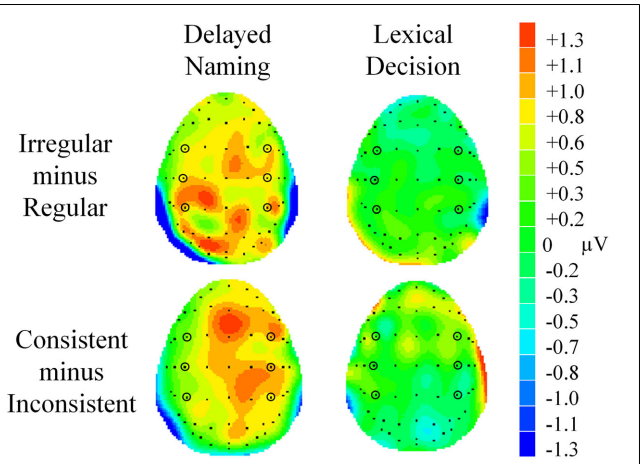


FIGURE 4 | Voltage maps showing effects of regularity and consistency in delayed naming and lexical decision in P200. The maps were calculated from the mean difference values (in μV) of irregular minus regular characters and consistent minus inconsistent characters between 200–270 ms. Electrodes FC3, FC4, C3, C4, CP3, and CP4 are circled for reference.

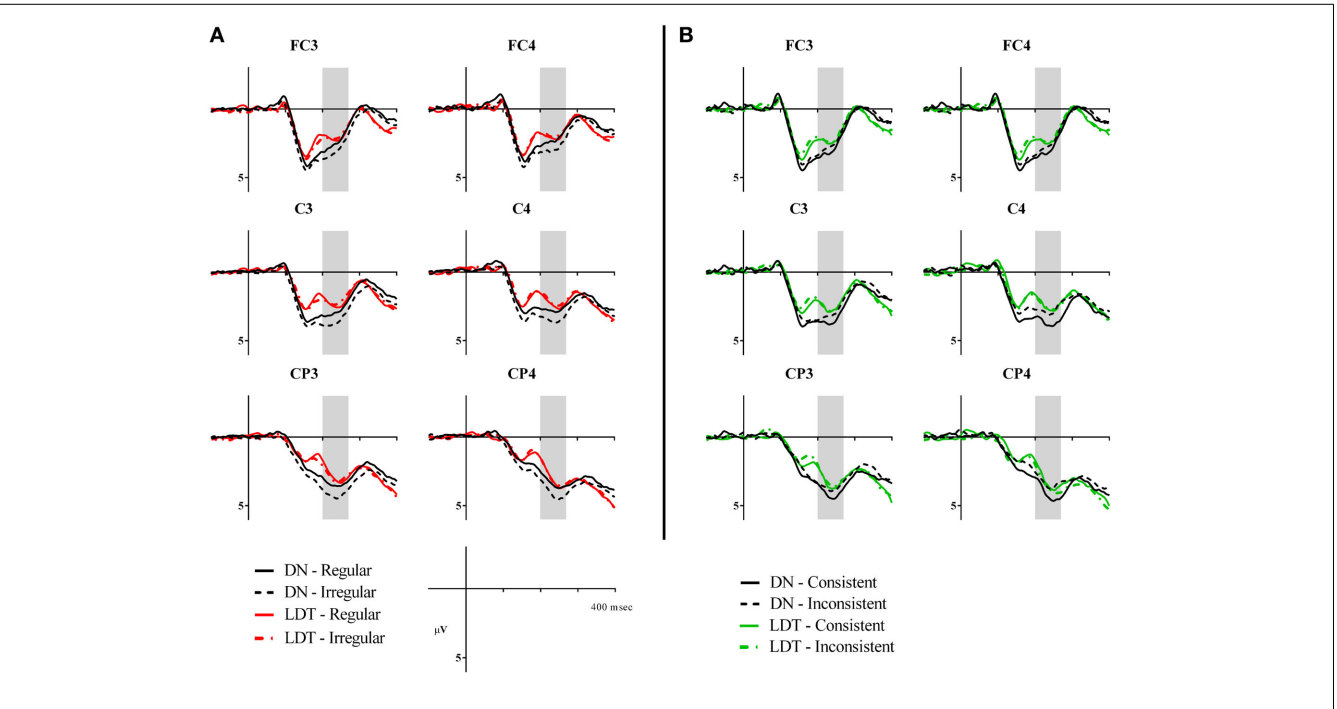


FIGURE 3 | (A) Grand average waveforms showing the P200 regularity effect at central sites (FC3, FC4, C3, C4, CP3, and CP4). **(B)** Grand average waveforms showing the P200 consistency effect at the same electrodes. Shaded areas represent the analysis window of 200–270 ms.

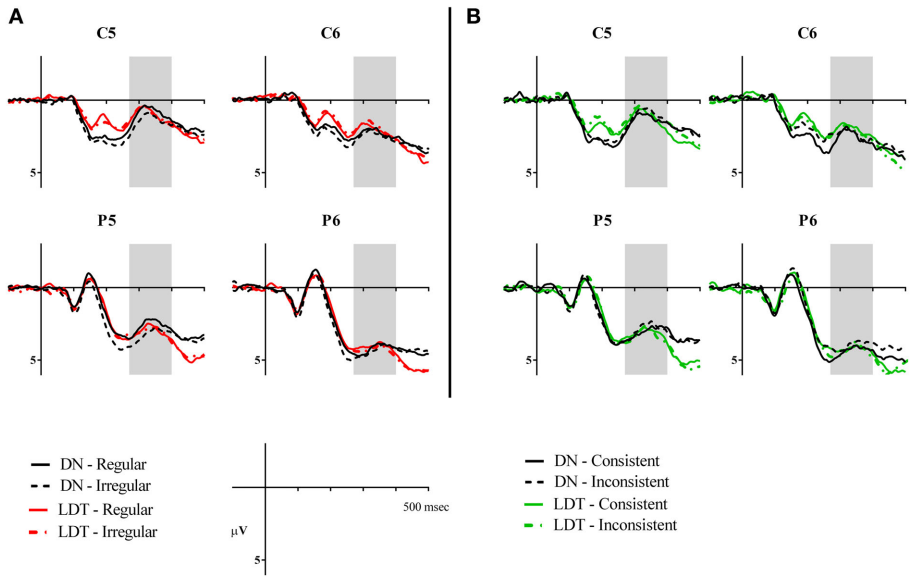


FIGURE 5 | (A) Grand average waveforms showing the N400 regularity effect at central and parietal sites (C5, C6, P5, and P6). **(B)** Grand average waveforms showing the N400 consistency effect at the same electrodes. Shaded areas represent the analysis window of 270–400 ms.

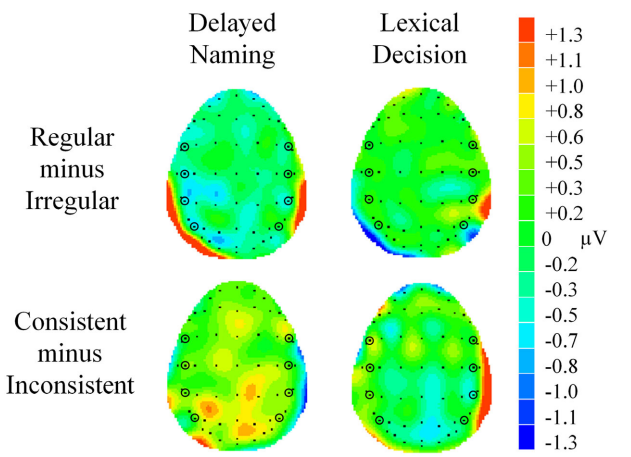


FIGURE 6 | Voltage maps showing effects of regularity and consistency in delayed naming and lexical decision in N400. The maps were calculated from the mean difference values (in μV) of regular minus irregular characters and consistent minus inconsistent characters between 270–400 ms. Electrodes FC5, FC6, C5, C6, CP5, CP6, P5, and P6 are circled for reference.

using behavioral and ERP measures, and how access to phonological information may be affected by task demands employing LD and DN tasks. It differed from previous reports in that both ortho-phonological effects were studied using ERPs and the patterns of these effects were contrasted between a task explicitly requiring phonological access and one without.

While the main foci of this study were on effects of regularity and consistency and their manifestation as a function of task, the results of lexicality effects in LD would ensure that the participants engaged in the task and the observations of the phonological effects from that task were reliable and valid. Our participants responded to real characters more quickly and accurately than pseudo-characters (546 ms and 97% vs. 661 ms and 89%, respectively). Moreover, they exhibited the typical pattern of greater negativity in N400 to pseudo- than real characters (Bentin et al., 1985; Holcomb, 1993; Nobre and McCarthy, 1994).

Although the behavioral measures concerning regularity and consistency effects were not significant, the ERP results clearly demonstrated the independence of regularity and consistency in terms of different time courses and directions of the effects in specific ERP components, mainly in DN. The effect of regularity, as predicted, was evident very early on in the N170 time window, followed by P200 and N400. Compared with irregular characters, regular characters exhibited more negative N170, less positive P200, and more negative N400. On the contrary, the consistency contrast was reliable only in P200, importantly with consistent characters eliciting more positive response than inconsistent ones. These findings reflect that the two effects have distinct neural correlates. The interaction of these effects with task in N170 and P200, with null effects of regularity and consistency in LD, indicates that phonological information is not automatically accessed during character recognition. Our observations differed most notably from previous work on the consistency effect on character naming in that it was detected only in P200 and was stronger for consistent characters, compared with the presence of the effect in N170, P200, and N400, and more positive P200 for inconsistent characters as reported in Hsu et al. (2009). We have questioned earlier whether consistency in previous ERP studies was confounded with regularity as the characters of the "consistent" condition had a consistency value very close to 1. If the consistency contrast in those studies was indeed one of regularity, i.e., consistent being equivalent to regular and inconsistent

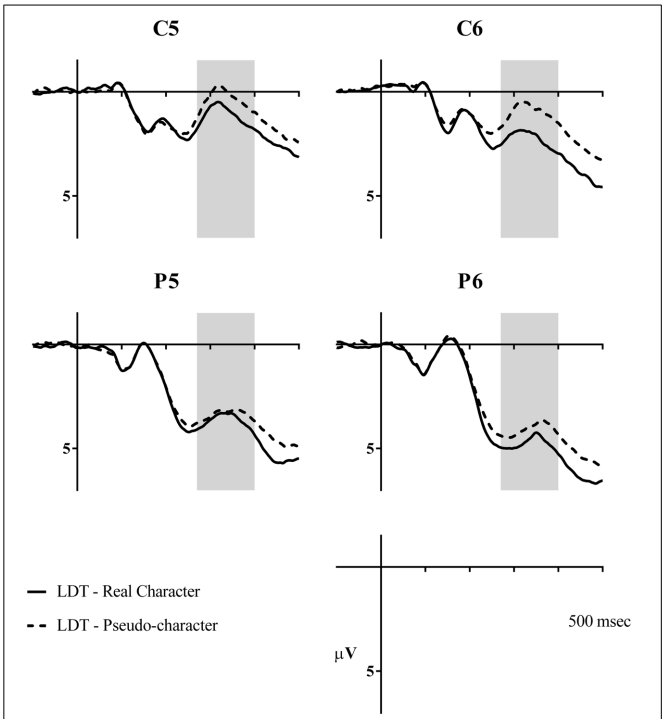


FIGURE 7 | Grand average waveforms showing the N400 lexicality effect at central and parietal sites (C5, C6, P5, and P6). Shaded areas represent the analysis window of 270–400 ms.

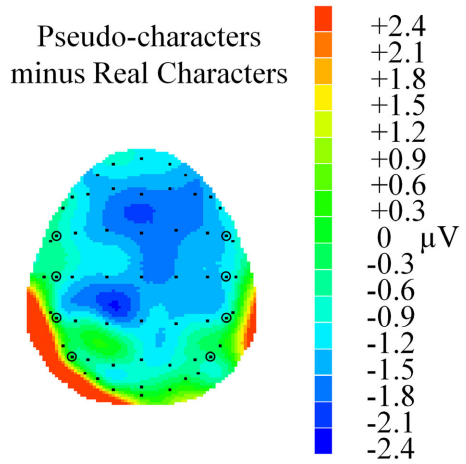


FIGURE 8 | Voltage map showing the lexicality effect in the lexical decision task in N400. The map was calculated from the mean difference (in μV) of pseudo-characters minus real characters between 270–400 ms. Electrodes FC5, FC6, C5, C6, CP5, CP6, P5, and P6 are circled for reference.

to irregular, then the present findings of regularity effects have exactly the same time course and pattern across the three ERP components as "consistency" effects in Hsu et al. (2009) with phonograms of large family size and with neighborhood characteristics well controlled for.

The regularity effect emerged within 200 ms post-stimulus onset. Its occurrence reflects the presence of conflict between phonological forms in irregular phonograms, i.e., those of the phonogram and its phonetic radical. The effect in N170 entails orthographic analysis of the phonogram into its radicals and mapping from the character and stand-alone radical(s) to phonology. This ERP component has been associated with identification of radicals (e.g., Hsiao et al., 2007; Su et al., 2012). Phonological modulations during character recognition on N170 have also been documented, although described as an effect of consistency (Lee et al., 2007; Hsu et al., 2009). More negative N170 for regular characters may reflect greater activation of a single phonological form or facility in processing because of an absence of conflict. The following component, P200, exhibited greater positivity for irregular characters. The direction of the contrast between regular and irregular characters is compatible with that of "consistency" effect in Hsu et al. (2009), as well as regularity effects in English (e.g., Sereno et al., 1998). Stronger P200 may be interpreted as more effortful processing due to competition between two phonological forms. Finally, the observation of more negative N400 for regular than irregular phonograms complements the findings in Lee et al. (2006b) of earlier onset and longer duration of N400 semantic priming effects elicited by regular phonograms. The present result can likewise be interpreted as interaction between phonological and semantic information in this time window. It reflects greater processing effort when different word meanings are mapped onto the same phonological form. One may argue that this account would be relevant to a task that explicitly accesses phonology, i.e., DN, but not necessarily LD. Although the interaction between task and regularity did not reach significance ($p = 0.101$) and only a main effect of regularity was found, inspection of **Figures 5**, **6** reveals a tendency of greater negativity for regular than irregular characters in DN, particularly for left hemisphere electrodes, but minimal difference in LD.

Reliable effects of consistency were only obtained in P200. Their later appearance, compared with regularity effects, can be explained in terms of competition among orthographic neighbors induced by the phonetic radical of the target phonogram. The identification of the phonetic radical, which takes place during the N170 time window, must precede the activation of phonograms sharing that radical. The shorter duration of the consistency effect may be attributed to the fact that activation of the orthographic neighbors is not sustained by orthographic forms in the stimulus, and the assumption that further access to semantics by activated non-target phonograms is irrelevant to a naming task. The opposite effects of regularity and consistency on P200 have provided critical evidence for their distinction. The result seems counterintuitive in that consistent characters showed greater P200 than inconsistent ones. Note that the contrast between consistent and inconsistent phonograms in this study was a matter of degree. To distinguish consistency from regularity, we included an equal number of regular and irregular characters for the consistent and inconsistent conditions. The consistency values by type or token (**Table 1**) of phonograms in the "consistent" condition were far from 1, differing from previous investigations. The stimuli in the two consistency conditions were matched in number

of stroke, character frequency, orthographic and phonological neighborhood sizes. However, as one would expect, inconsistent characters had more phonological alternatives than consistent phonograms. Following the reasoning of greater competition revealed in stronger P200 in the case of regularity, we propose that fewer phonological competitors actually induce stronger inhibition among one another than when there are more competitors. In the former situation, each phonological form is activated by a larger number of phonograms, while in the latter, activation or competition is more distributed, resulting in weaker mutual interference. Our account is contrary to the one in Lee et al. (2006a) where greater P200 was found for inconsistent characters because more phonological candidates were activated initially. We have tried to argue in this paper that consistency was conflated with regularity in Lee et al. (2006a, 2007) and Hsu et al. (2009), and our findings of regularity parallel theirs of consistency in P200.

The present findings, together with those reported in the works by Lee and colleagues, have clearly demonstrated that skilled readers of Chinese can access phonology from characters within 200 ms in a reading task. This observation differs dramatically from studies of alphabetic scripts, which have traditionally focused on late components. They include the N400 and the following late positive complex (LPC), which occurs between 500 and 800 ms over the left centro-parietal region and is generally interpreted as reflecting conflict resolution and word recognition memory (see Rugg and Curran, 2007; Van Petten and Luka, 2012 for review). The interest in late components might be due to the fact that in the alphabetic writing system whole-word phonology is only available upon or after lexical access, which is believed to take place at the N400 time window (see Lau et al., 2008; Kutas and Federmeier, 2011 for review). However, as argued in Sereno and Rayner (2003), reading research employing the eye movement method has consistently found that the average fixation duration of normal adult readers is around 250 ms. Hence, the focus on N400 and LPC is unlikely to reveal the full picture of online stages of processing during word recognition.

Ample evidence has been accumulated to establish the N170 as an index of one's sensitivity to print (see Maurer and McCandliss, 2007 for review). The reading-related N170 is obtained when words, pseudowords, and letter strings are contrasted with nonlinguistic visual forms, and left lateralization of the component is indicative of reading expertise (Maurer et al., 2008). Maurer and McCandliss (2007) have further put forth the phonological mapping hypothesis that the degree of lateralization of N170 may be correlated with the depth of an orthographic system. To illustrate, readers of German, a transparent script, showed comparable left-lateralized N170 to real words and pseudowords, while readers of English, a more opaque script, exhibited stronger effects for words than pseudowords (Maurer et al., 2005). Maurer and colleagues proposed that the left lateralization of N170 is related to the exposure to grapheme-phoneme conversion during reading acquisition. Hence, one would not expect a left-lateralized N170 in readers of a logographic system such as Chinese. The prediction seems to find support from Kim et al. (2004) in which native Korean speakers learning English and Chinese as second languages (L2) were presented with words in these languages as well as pictures in a semantic categorization task. Left-lateralized N170 responses were observed for Korean and English, while bilateral distribution of the component was seen for Chinese and pictures. However, as little information was provided for the participants' proficiency in English and Chinese, it is not clear whether differential responses were due to the properties or the levels of proficiencies of their L2s. In fact, Maurer et al. (2008) found stronger N170 in the left hemisphere of Japanese native speakers to all three Japanese scripts, i.e., *katakana*, *hiragana*, *kanji* characters, compared to their less familiar English script. In studies involving native Chinese readers, the results are mixed with respect to the laterality of N170. Lee et al. (2007), Hsu et al. (2009), as well as the present study, did not find hemispheric dominance of the component; nonetheless, left-lateralization of the N170 has been obtained in adults (Lin et al., 2011; Zhao et al., 2012) and children (Cao et al., 2011; Su et al., 2013). In sum, the N170 has consistently been associated with skilled reading in different writing systems, and left-lateralization of the component is not necessarily influenced by the nature of orthography-phonology mapping.

ERP studies of alphabetic scripts showing effects of GPC on P200 and N400 are equally few. Sereno et al. (1998) obtained greater P200 peaking at 168 ms post-stimulus in the centro-frontal region for low frequency regular than exception English words in a LD task. However, the effect was observed from a subset, 13 out of 32, of the participants. The interaction between phonology and orthography reflected in N400 seems to be restricted to pseudohomophones, i.e., pseudowords that sound like real words. Briesemeister et al. (2009) found weaker effects of pseudohomophones in contrast with pseudoword controls in N400 in LD of German words. Comparable facilitative effects in N400 were shown for pseudohomophones and semantically related words, compared with semantically incongruent words and pseudowords, in the context of semantically constrained sentences (Newman and Connolly, 2004). Although results of these two studies suggested that phonological information generated from print played a role in reading during the N400 time window, they did not come from a direct contrast of regularity or consistency.

The influence of orthography-phonology mapping on word reading in Chinese and the alphabetic writing system is shown to be very different, as evidenced by the manifestations of the effects from early to late ERP components. In Chinese, the regularity effect occurs simultaneously as orthographic analysis begins to take place and persists until lexical recognition; its emergence is immediately followed by the relatively short-lived consistency effect. On the other hand, little evidence points to early occurrence of these effects in alphabetic scripts. This contrast seems counterintuitive at first glance. We propose that the different time courses of access to phonology from print stems from a fundamental difference in the nature of orthography-phonology mapping between the two orthographic systems, namely, addressed phonology in the logographic system and inherently assembled phonology in alphabetic scripts. As described in the Introduction and argued in Law et al. (2009), phonological access is always lexical in Chinese word processing. In the contrast of regularity, the competition is between the pronunciation of the phonogram

and that of its phonetic radical as a character. In the case of consistency, the competition is among phonological forms of phonograms sharing the same phonetic radical, regardless of the lexical status of the phonetic radical. Our view differs from the two-stage framework consisting of sublexical and lexical processing indexed by the P200 and N400, respectively (Lee et al., 2007). However, we believe that the divergence is superficial and a matter of wording, since the underlying mechanism of character naming portrayed in Lee et al. (2007) is essentially the same as the one presented here.

In conclusion, the main findings of this investigation, along with those of previous ERP studies of Chinese character reading, have captured a basic difference between logographic and alphabetic writing systems in terms of phonological access from visual word in the early stages of lexical recognition. The conceptual distinction between regularity and consistency in Chinese allowed us to examine their effects independently. The different mechanisms were clarified through their occurrence across ERP components. Finally, the comparison between LD and DN has demonstrated that access to phonological information from print is not automatic and subject to task demands.

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: Yen Na Yum, Sam-Po Law, I-Fan Su, and Kai-Yan Dustin Lau. Collected and analyzed the data: Yen Na Yum and Kwan Nok Mo. Interpreted the data: Yen Na Yum, Sam-Po Law, I-Fan Su, Kai-Yan Dustin Lau, and Kwan Nok Mo. Wrote the paper: Yen Na Yum and Sam-Po Law.

ACKNOWLEDGMENTS

This research was supported by a Faculty Research Fund at the Faculty of Education, the University of Hong Kong (Project titled "Sublexical Processing in First and Second Language Chinese Users").

REFERENCES

Andrews, S. (1982). Phonological recoding: is the regularity effect consistent? *Mem. Cognit.* 10, 565–575. doi: 10.3758/BF03202439

Baron, J., and Strawson, C. (1976). Use of orthographic and word-specific knowledge in reading words aloud. *J. Exp. Psychol. Hum. Percept. Perform.* 2, 386–393. doi: 10.1037/0096-1523.2.3.386

Bentin, S., McCarthy, G., and Wood, C. C. (1985). Event-related potentials, lexical decision, and semantic priming. *Electroencephalogr. Clin. Neurophysiol.* 60, 343–355. doi: 10.1016/0013-4694(85)90008-2

Briesemeister, B. B., Hofmann, M. J., Tamm, S., Kuchinke, L., Braun, M., and Jacobs, A. M. (2009). The pseudohomophone effect: evidence for an orthography–phonology- conflict. *Neurosci. Lett.* 455, 124–128. doi: 10.1016/j.neulet.2009.03.010

Cao, X., Li, S., Zhao, J., and Weng, X. C. (2011). Left-lateralized early neurophysiological response for Chinese characters in young primary school children. *Neurosci. Lett.* 492, 165–169. doi: 10.1016/j.neulet.2011.02.002

Coltheart, M. (1978). "Lexical access in simple reading tasks," in *Strategies in Information Processing*, ed G. Underwood (London: Academic Press), 151–216.

Coltheart, M., Curtis, A., Atkins, B., and Haller, M. (1993). Models of reading aloud: dual route and parallel-distributed processing approaches. *Psychol. Rev.* 100, 589–608. doi: 10.1037/0033-295X.100.4.589

Coltheart, M., Rastle, K., Perry, C., Langdon, R., and Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychol. Rev.* 108, 204–256. doi: 10.1037/0033-295X.108.1.204

Cortese, M. J., and Simpson, G. B. (2000). Regularity effects in word naming: where are they? *Mem. Cognit.* 28, 1269–1276. doi: 10.3758/BF03211827

Ding, G., Peng, D., and Taft, M. (2004). The nature of the mental representation of radicals in Chinese: a priming study. *J. Exp. Psychol. Learn. Mem. Cogn.* 30, 530–539. doi: 10.1037/0278-7393.30.2.530

Fang, S.-P., Horng, R.-Y., and Tzeng, O. J. L. (1986). "Consistency effects in the Chinese characters and pseudo-character naming tasks," in *Linguistics, Psychology, and the Chinese Language*, eds H. S. R. Kao and R. Hoosain (Hong Kong: Centre of Asian Studies, University of Hong Kong), 11–21.

Glushko, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. *J. Exp. Psychol. Hum. Percept. Perform.* 5, 574–691. doi: 10.1037/0096-1523.5.4.674

Gough, P. B., and Cosky, M. J. (1977). "One second of reading again," in *Cognitive Theory*, Vol. 2, eds N. J. Castellan, D. B. Pisoni, and G. R. Potts (Hillsdale, NJ: Erlbaum), 271–288.

Hillis, A., and Caramazza, A. (1995). Converging evidence for the interaction of semantic and phonological information in accessing lexical information for spoken output. *Cogn. Neuropsychol.* 12, 187–227. doi: 10.1080/02643299508251996

Holcomb, P. J. (1993). Semantic priming and stimulus degradation: implications for the role of the N400 in language processing. *Psychophysiology* 30, 47–61. doi: 10.1111/j.1469-8986.1993.tb03204.x

Hsiao, J. H. W., Shillcock, R., and Lee, C. C. Y. (2007). Neural correlates of foveal splitting in reading: evidence from an ERP study of Chinese character recognition. *Neuropsychologia* 45, 280–1292. doi: 10.1016/j.neuropsychologia.2006.10.001

Hsu, C. H., Tsai, J. L., Lee, C. Y., and Tzeng, O. J. L. (2009). Orthographic combinability and phonological consistency effects in reading Chinese phonograms: an event-related potential study. *Brain Lang.* 108, 56–66. doi: 10.1016/j.bandl.2008.09.002

Hue, C. W. (1992). "Recognition processes in character naming," in *Language Processing in Chinese*, eds H.-C. Chen and O. J. L. Tzeng (Amsterdam: North-Holland), 93–107. doi: 10.1016/S0166-4115(08)61888-9

Jared, D. (2002). Spelling-sound consistency and regularity effects in word naming. *J. Mem. Lang.* 46, 723–750. doi: 10.1006/jmla.2001.2827

Kay, J., and Bishop, D. (1987). "Anatomical differences between nose, palm, and foot, or, the body in question: further dissection of the processes of sub-lexical spelling-sound translation," in *Attention and Performance XII: The Psychology of Reading*, ed M. Coltheart (Hillsdale, NJ: Erlbaum), 449–469.

Kim, K. H., Yoon, H. W., and Park, H. W. (2004). Spatiotemporal brain activation pattern during word/picture perception by native Koreans. *Neuroreport* 15, 1099–1103. doi: 10.1016/j.bandl.2008.09.002

Kutas, M., and Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annu. Rev. Psychol.* 62, 621–647. doi: 10.1146/annurev.psych.093008.131123

Lau, E. F., Phillips, C., and Poeppel, D. (2008). A cortical network for semantics: (De)constructing the N400. *Nat. Rev. Neurosci.* 9, 920–933. doi: 10.1038/nrn2532

Law, S.-P., Weekes, B. S., Wong, W., and Chiu, K. (2009). Reading aloud pseudo-characters by individuals with acquired dyslexia: evidence for lexically-mediated processes in reading Chinese. *Lang. Cogn. Process.* 24, 983–1008. doi: 10.1080/01690960802193696

Lee, C.-Y. (2008). Rethinking of the regularity and consistency effects in reading. *Lang. Linguist.* 9, 177–186.

Lee, C. Y., Tsai, J. L., Chan, W. H., Hsu, C. H., Hung, D. L., and Tzeng, O. J. L. (2007). Temporal dynamics of the consistency effect in reading Chinese: an event-related potentials study. *Neuroreport* 18, 47–151. doi: 10.1097/WNR.0b013e328010d4e4

Lee, C. Y., Tsai, J. L., Chiu, Y. C., Tzeng, O. J. L., and Hung, D. L. (2006a). The early extraction of sublexical phonology in reading Chinese pseudocharacters: an event-related potentials study. *Lang. Linguist.* 7, 619–636.

Lee, C. Y., Tsai, J. L., Huang, H. W., Hung, D. L., and Tzeng, O. J. L. (2006b). The temporal signatures of semantic and phonological activations for Chinese sublexical processing: an event-related potential study. *Brain Res.* 1121, 150–159. doi: 10.1016/j.brainres.2006.08.117

Lee, C.-Y., Tsai, J.-L., Su, E. C.-I., Tzeng, O. J. L., and Hung, D. L. (2005). Consistency, regularity, and frequency effects in naming Chinese characters. *Lang. Linguist.* 6, 75–107.

Li, Z.-M. (1989). *Lishi Zhongwen Zidian [Li Dictionary of Chinese characters]*. Hong Kong: Chinese University Press.

Lian, Y.-W. (1985). *Zhongwen Nianzi Licheng de Tantao: Shengpang de Yuyin Chufa Zuoyong [An Investigation Of The Processes In Character Naming in Chinese: The Influence of the Phonetic Component]*. Unpublished master's thesis, National University of Taiwan, Taipei.

Lin, S. E., Chen, H. C., Zhai, J., Li, S., He, S., and Weng, X. C. (2011). Left-lateralized N170 response to unpronounceable pseudo but not false Chinese characters - the key role of orthography. *Neurosci* 190, 200–206. doi: 10.1016/j.neuroscience.2011.05.071

Maurer, U., Brandeis, D., and McCandliss, B. D. (2005). Fast, visual specialization for reading in English revealed by the topography of the N170 ERP response. *Behav. Brain Funct.* 1:13. doi: 10.1186/1744-9081-1-13

Maurer, U., and McCandliss, B. D. (2007). "The development of visual expertise for words: the contribution of electrophysiology," in *Single-Word Reading: Behavioral and Biological Perspectives*, eds E. L. Grigorenko and A. Naples (Mahwah, NJ: Lawrence Erlbaum Associates), 43–64. doi: 10.1080/87565641.2010.480916

Maurer, U., Zevin, J. D., and McCandliss, B. D. (2008). Left-lateralized N170 effects of visual expertise in reading: evidence from Japanese syllabic and logographic scripts. *J. Cogn. Neurosci.* 20, 1878–1891. doi: 10.1162/jocn.2008.20125

Newman, R. L., and Connolly, J. F. (2004). Determining the role of phonology in silent reading using event-related brain potentials. *Cogn. Brain Res.* 21, 94–105. doi: 10.1016/j.cogbrainres.2004.05.006

Ni, H.-S. (1982). *Xiandai Hanzi Xingshengzi Zihui [A Dictionary of Contemporary Chinese Phonetic Compounds]*. Beijing: Yuwen Chubanshe.

Nobre, A. C., and McCarthy, G. (1994). Language-related ERPs: scalp distributions and modulation by word type and semantic priming. *J. Cog. Neurosci.* 6, 233–255. doi: 10.1162/jocn.1994.6.3.233

Peereman, R. (1995). Naming regular and exception words: further examination of the effect of phonological dissension among lexical neighbours. *Eur. J. Cogn. Psychol.* 7, 307–330. doi: 10.1080/09541449508402451

Perfetti, C. A., Liu, Y., and Tan, L. H. (2005). The lexical constituency model: some implications of research on chinese for general theories of reading. *Psychol. Rev.* 112, 43–59. doi: 10.1037/0033-295X.112.1.43

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., and Patterson, K. E. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychol. Rev.* 103, 56–115. doi: 10.1037/0033-295X.103.1.56

Rugg, M. D., and Curran, T. (2007). Event-related potentials and recognition memory. *Trends Cogn. Sci.* 11, 251–257. doi: 10.1016/j.tics.2007.04.004

Seidenberg, M. S. (1985). The time course of phonological code activation in two writing systems. *Cognition* 19, 1–30. doi: 10.1016/0010-0277(85)90029-0

Seidenberg, M. S., and McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychol. Rev.* 96, 523–568. doi: 10.1037/0033-295X.96.4.523

Seidenberg, M. S., Waters, G. S., Barnes, M. A., and Tanenhaus, M. K. (1984). When does irregular spelling or pronunciation influence word recognition? *J. Verbal Learn. Verbal Behav.* 23, 383–404. doi: 10.1016/S0022-5371(84)90270-6

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

Sereno, S. C., Rayner, K., and Posner, M. I. (1998). Establishing a time-line of word recognition: evidence from eye movements and event-related potentials. *Neuroreport* 9, 2195–2200. doi: 10.1097/00001756-199807130-00009

Stanovich, K., and Bauer, D. (1978). Experiments on the spelling-to-sound regularity effect in word recognition. *Mem. Cognit.* 6, 410–415. doi: 10.3758/BF03197473

Su, I.-F., Lau, D., and Law, S.-P. (2013). Neural correlates of normal reading development and reading disorders in Chinese: preliminary findings from event-related potentials. *Procedia Soc. Behav. Sci.* 94, 187–188. doi: 10.1016/j.sbspro.2013.09.092

Su, I. S., Mak, S. C., Ching, L. Y., and Law, S. P. (2012). Taking a radical position: evidence for position specific radical representations in Chinese character recognition using masked priming ERP. *Front. Psychol.* 3:333. doi: 10.3389/fpsyg.2012.00333

Taft, M., and Zhu, X. (1997). Sub-morphemic processing in reading Chinese. *J. Exp. Psychol. Learn. Mem. Cogn.* 23, 761–775. doi: 10.1037/0278-7393.23.3.761

Van Petten, C., and Luka, B. J. (2012). Prediction during language comprehension: benefits, costs, and ERP components. *Int. J. Psychophysiol.* 83, 176–190. doi: 10.1016/j.ijpsycho.2011.09.015

Waters, G. S., Seidenberg, M. S., and Bruck, M. (1984). Children's and adults' use of spelling sound information in three reading tasks. *Mem. Cognit.* 12, 293–305. doi: 10.3758/BF03197678

Weekes, B., Chen, M. J., and Lin, Y.-B. (1998). "Differential effects of phonological priming on Chinese character recognition," in *Cognitive Processing of the Chinese and the Japanese Languages*, eds C. K. Leong and K. Tamaoka (Netherlands: Kluwer Academic Publishers), 47–68. doi: 10.1007/978-94-015-9161-4_3

Wu, J. T., Chou, T. L., and Liu, I. M. (1994). "Zhongwen zici chuli guocheng li de pinlu xiaoguo fenxi," [The frequency effects in processing Chinese characters and words] in *Advances in the Study of Chinese Language Processing*, eds H. W. Chang, J. T. Huang, C. W. Hue, and O. J. L. Tzeng (Taipei: Taiwan University), 31–57.

Zhao, J., Li, S., Lin, S. E., Cao, X. H., He, S., and Weng, X. C. (2012). Selectivity of N170 in the left hemisphere as an electrophysiological marker for expertise in reading Chinese. *Neurosci. Bull.* 28, 577–584. doi: 10.1007/s12264-012-1274-y

Zhou, X., and Marslen-Wilson, W. (1999). "Sublexical processing in reading Chinese," in *Reading Chinese Script: A Cognitive Analysis*, eds J. Wang, A. Inhoff, and H.-C. Chen (Hillsdale, NJ: Lawrence Erlbaum Associates), 37–63.

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.

Received: 21 February 2014; accepted: 26 March 2014; published online: 14 April 2014.
Citation: Yum YN, Law S-P, Su I-F, Lau K-YD and Mo KN (2014) An ERP study of effects of regularity and consistency in delayed naming and lexicality judgment in a logographic writing system. Front. Psychol. ***5****:315. doi: 10.3389/fpsyg.2014.00315*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 15 July 2014
doi: 10.3389/fpsyg.2014.00708

Measuring inconsistencies can lead you forward: Imageability and the x-ception theory

Sara Dellantonio[1]*, Claudio Mulatti[2], Luigi Pastore[3] and Remo Job[1]

[1] *Psychology and Cognitive Science, Università degli Studi di Trento, Trento, Italy*
[2] *Department of Developmental Psychology and Socialisation, Università degli Studi di Padova, Padova, Italy*
[3] *Department of Educational Sciences, Psychology, Communication, Università degli Studi di Bari, Bari, Italy*

Edited by:
Davide Crepaldi, University of Milano-Bicocca, Italy

Reviewed by:
Lotte Meteyard, University of Reading, UK
Christos Pliatsikas, University of Kent, UK

****Correspondence:***
Sara Dellantonio, Dipartimento di Psicologia e Scienze Cognitive, Università degli Studi di Trento, Corso Bettini, 84, 38068 Rovereto, Italy
e-mail: sara.dellantonio@unitn.it

According to the traditional view, both imageability and concreteness ratings reflect the way word meanings rely on information mediated by the senses. As a consequence, the two measures should and do correlate. The link between these two indexes was already hypothesized and demonstrated by Paivio et al. (1968) in a seminal article, where they introduced the idea of imageability ratings for the first time. However, in this first study, they also noted a contrasting pattern in the ratings for imageability and concreteness with some words that refer to affective attitudes or emotional states receiving high imageability but low concreteness ratings. Recent studies confirm this inconsistency (e.g., Altarriba and Bauer, 2004) leading to the claim that emotion words form a particular class of terms different from both concrete and abstract words. Here we use the MRC psycholinguistic database to show that the there are other classes of terms for which imageability and concreteness are uncorrelated. We show that the common feature of these word classes is that they directly or indirectly refer to proprioceptive, interoceptive, or affective states, i.e., to internal, body-related, sensory experiences. Thus, imageability and concreteness can no longer be considered interchangeable constructs; rather, imageability is a different, and perhaps more interesting, measure: it not only reflects the ease with which memories of external events come to mind, as previously hypothesized, but also reflects the ease with which memories of internal events come to mind.

Keywords: imageability, concreteness, abstract words, concrete words, emotion words

INTRODUCTION

Analogously to frequency and familiarity, concreteness, and imageability are properties of words (referents) that are partially intertwined. According to the traditional view, both imageability and concreteness ratings reflect the way word meanings rely on information mediated by the senses. For this reason, they are the most relevant operational constructs to address the question of the processing differences between concrete and abstract words, i.e., between words denoting things that can be perceived by the senses and words that do not have this kind of reference. Since they are hypothesized to detect analogous properties, the two measures should strongly correlate, as indeed they do. And because of this strong correlation, in experiments for the selection of concrete vs. abstract verbal material they are often used interchangeably (see e.g., Reilly and Kean, 2007; Connell and Lynott, 2012).

The link between these two indexes was already hypothesized and demonstrated by Paivio et al. (1968), who introduced the idea of imageability ratings for the first time. However, already in this study a contrasting pattern was reported and has not yet been fully accounted for: a number of words referring mainly to affective attitudes or emotional states received high imageability ratings but low concreteness ratings. Recent studies confirm this anomaly (e.g., Altarriba et al., 1999; Wiemer-Hastings et al., 2001; Altarriba and Bauer, 2004; Wiemer-Hastings and Xu, 2005) and maintain that emotion words form a particular class of terms different from both concrete and abstract terms.

In this paper we present an account of this presumed inconsistency arguing that imageability ratings measure not only whether (how much) words rely on external sensory information, but also whether (how much) words rely on *internal bodily-related sensory experience*. Since imageability ratings are a joint measure of the link between word meanings and both external and internal sensory experience, while concreteness ratings measure the link to external sensory information only, we suggest that an index of the "weight" internal sensory information has with respect to the meaning of a word can be obtained by subtracting the concreteness rating of this word from its imageability rating.

To support this view we use the MRC psycholinguistic database (Coltheart, 1981; Wilson, 1988). First of all, on the basis of an analysis of the instructions used when collecting the imageability ratings included in the database, we suggest that the contrasting pattern observed with respect to these word classes is due to the fact that imageability ratings do not only reflect the ease/difficulty with which people can evoke a mental picture of the instances denoted by the word, as it commonly assumed (e.g., Vigliocco et al., 2009; Connell and Lynott, 2012). These instructions might rather have biased people to assign their imageability ratings on

the basis of the ease/difficulty with which a word arouses sensory experience of any kind, including internal, body-related sensations. Secondly, we analyze the imageability and concreteness ratings of specific words included in these databases to show that, in addition to emotion words, there are also other classes of terms for which imageability and concreteness are uncorrelated and that the common feature of all kinds of words exhibiting this contrasting pattern is that they directly or indirectly refer to proprioceptive, interoceptive, or affective states, i.e., to internal, body-related, sensory experiences.

This shows that imageability and concreteness are not interchangeable constructs: imagery ratings are not only another means to assess the degree of concreteness of a word, but they are also a different, and perhaps more interesting, measure of the link between a word and some internal information pertaining to the state the word denotes. Since imageability ratings are a joint measure of the connections between words and both external and internal sensory experience, subtracting concreteness from imageability gives us a tangible measure of the internal sensory information aroused by a word. Even though this measure cannot be completely accurate since it is a result of ambiguous norming instructions, it can still indicate whether a word relies (more or less heavily) on body-related sensory information. Clearly, new collections of ratings on the basis of less ambiguous instructions are required in order to have more precise imageability ratings to use for experimentation. However, our study indicates a way to interpret the imageability construct from a new and possibly more fruitful perspective which allows us to both avoid the incongruities of the old measure and to assess more clearly in what respects concreteness and imageability converge and in what respects they instead diverge.

IMAGEABILITY IN THE MRC PSYCHOLINGUISTIC DATABASE

Concrete (CONC) ratings for the MRC psycholinguistic databases were originally collected by Spreen and Schulz (1966). With the exception of the most recent database of 2013 (Brysbaert et al., 2013), later collections are included in the MRC database and rely on the same definition of concreteness and on the same instructions. As Spreen and Schulz (1966, p. 459) point out, starting from the twenties it became clear that there are differences in remembering, recognizing and understanding concrete and abstract verbal material. For the study of these differences they tried to work out a precise scale for concrete and abstract words, based on a non-ambiguous definition of the two poles. In fact, previous definitions interpreted the opposition of abstractness and concreteness in at least two different ways: on the one hand, in terms of general, i.e., generic vs. specific, and on the other, in terms of a difference in the nature of the referents of abstract and concrete terms–the referents of concrete terms are directly connected to sensory experience, while those of abstract words lack this connection. Spreen and Schulz (1966) opted for this last definition and suggested that the scale should measure whether the referents of a word can or cannot be experienced by the senses.

The imageability (IMAG) scale was first introduced by Paivio et al. (1968) as a further measure in addition to CONC to investigate psychological effects of linguistic abstractness-concreteness: in the context of their study, imagery is postulated to be "the major effective psychological attribute underlying abstractness-concreteness" (p. 2). High and low imagery ratings measure the ease or difficulty with which words arouse sensory images. These sensory images are defined in a rather vague manner as any kind of sensory experience evoked by words by recalling non-verbal representations of their referent. Concreteness is considered to be determined independently from imagery; however, highly concrete words are assumed to have a high image-arousing value, since they are particularly effective in evoking sensory images of their referents which, in this case, consist mainly in mental pictures of them.

The correlation between IMAG and CONC was confirmed by Paivio et al. (1968) on 925 words. This correlation provided evidence for Paivio's *Dual Code Theory* (Paivio, 1971, 1986, 2007), according to which cognitive processing is carried out on the basis of two different subsystems, "one specialized for the representation and processing of information concerning nonverbal objects and events, the other specialized for dealing with language (Paivio, 1986, p. 53). The nonverbal (symbolic) subsystem is referred to as the imagery system: "its critical functions include the analysis of scenes and the generation of mental images (both functions encompassing other sensory modalities in addition to visual)" (Paivio, 1986, pp. 53–54). The language-specialized system is called the verbal system. In Paivio's perspective, the nonverbal system of imagery is activated primarily by concrete (i.e., perceivable) stimuli (Paivio, 1986, p. 68), therefore words with a high CONC rating should have high IMAG ratings, while words with a low CONC ratings (i.e., abstract words) should have low IMAG ratings. According to Paivio, this dual system can also account for the so-called "concreteness effect," i.e., the fact that concrete words are processed more easily and quickly than abstract words because, while abstract words activate verbal representations only, concrete words activate representations in both the verbal and in the imagery system, and this facilitates the referential act.

The idea that CONC and IMAG are strongly correlated both theoretically and from the point of view of their ratings has become established in the literature (Paivio et al., 1968 found a correlation of 0.83; this correlation has been confirmed using larger word numbers by several other studies that report values ranging from 0.64 to 0.95: see e.g., Christian et al., 1978; Toglia and Battig, 1978; Gilhooly and Logie, 1980; Rubin, 1980; Friendly et al., 1982; Rubin and Friendly, 1986; Schwanenflugel et al., 1988; Benjafield and Muckenheim, 1989). This is the reason why these two measures are considered interchangeable and are both used to study the processing differences between abstract and concrete verbal material. As e.g., Reilly and Kean (2007) point out: "Although imageability and concreteness are technically different psycholinguistic constructs, the correlation between these variables is so strong that many authors use the terms interchangeably. Here we make the same assumption of synonymy between imageability and concreteness in terms of theory (i.e., concreteness effects—imageability effects)" (p. 158). The same point has been made more recently by Connell and Lynott (2012): "Imageability ratings are frequently used interchangeably with concreteness ratings in the experimental literature [. . .] because

of their high correlation and theoretical relationship in dual coding theory" (p. 453).

However, even though the correlation between CONC and IMAG is quite strong, (a) it is lower than expected and (b) exhibits some relevant anomalies. As Paivio and colleagues underline: "the correlation between I and C, although substantial, is not as high as one might expect if it is assumed that both scales measure the same underlying variable" (Paivio et al., 1968, p. 7). According to Paivio et al. (1968) the problem is due to some sets of problematic items whose IMAG ratings are significantly greater than their CONC ratings[1]. As they note, these items exhibit an interesting semantic similarity: "Most of these are words with strong emotional and evaluative connotations. The largest group consists of terms referring to affective reactions or affective attitudes" (Paivio et al., 1968, p. 7).

Paivio et al. (1968) do not offer any explanation for the contrasting patterns for IMAG and CONC. However they suggest that also in these cases high IMAG must be due to the fact that the word easily evokes some kind of sensory experience, which in this case seems to be of an affective kind: "These words appear to have the common property of having been associated with sensory experience (usually affective in nature)" (p. 7). These observations open the door for the hypothesis also embraced by Paivio in his later work (1986, p. 79) that affective and emotional words have a high IMAG rating because they directly evoke the sensory experience of an affective arousal.

Similar inconsistencies are also pointed out by recent studies that interpret them in a way analogous to Paivio's suggesting that emotion words are different from both abstract and concrete words as regards their CONC and IMAG ratings and must therefore be considered as a particular word class with specific characteristics (Altarriba et al., 1999; Altarriba and Bauer, 2004; Wiemer-Hastings and Xu, 2005). In particular, as the results of Altrarriba's study (2004) indicate: "concepts represented by emotion words are more imageable and are easier to of a context for than abstract words but are less concrete than abstract words. They are less imageable, less concrete, and less likely to activate a context than concrete words" (p. 407). Therefore, even though emotion words are, as one would expect, less CONC than concrete words, they turn out to be also less CONC than abstract words, even though their IMAG is significantly greater than that of abstract words. Thus, for this word class the divergence between IMAG and CONC is particularly broad. As a matter of fact, our analysis of the IMAG and CONC ratings included in the MRC psycholinguistic database, which has been the main source for these measures showed that the difference between IMAG and CONC is significantly greater for emotion terms than for any other randomly chosen control group of words (we will come back to this aspect in the next section).

Altarriba et al. (1999), Altarriba and Bauer (2004) emphasize the fact that examining the unique qualities of emotion words with respect to other classes of terms is particularly important since it helps us understand how people recognize and label emotions. However, we think that the uniqueness of the IMAG and CONC ratings for emotion words can also help clarify the linguistic construct of imageability which is often considered vague and subject to different interpretations (e.g., Connell and Lynott, 2012; Westbury et al., 2013; Dellantonio et al., 2014). In fact, the anomaly of the IMAG and CONC ratings in the case of emotion words can be explained only by specifying what precisely IMAG measures and what is the specific difference between the constructs of IMAG and CONC. The key point to disentangle in this respect lies first of all in the content of the instructions given to subjects for assigning the CONC and the IMAG ratings included in the MRC database.

INSTRUCTION-BOUND RATINGS?

The original instructions for concreteness ratings were developed by Spreen and Schulz (1966), and then used in almost the same form by Paivio et al. (1968): however, while Spreen and Schulz (1966) labeled the end-points of the rating scales "low concreteness" and "high concreteness," Paivio et al. (1968) labeled them "high concreteness" and "high abstractness." Later collections used either the one or the other label interchangeably.

Spreen and Schulz's (1966) instructions for concreteness were: "Nouns may refer to persons, places, and things that can be seen, heard, felt, smelled, or tasted or to more abstract concepts that cannot be experienced by our senses. The purpose of this experiment is to rate a list of words with respect to "concreteness" in term of sense-experience. Any word that refers to objects, material or persons should receive a high concreteness rating; any word that refers to an abstract concept that cannot be experienced by the senses should receive a low concreteness rating. Think of the words "chair" and "independence." "Chair" can be experienced by our senses and therefore should be rated as high concrete; "independence" cannot be experienced by the senses as such and therefore should be rated as low concrete (abstract)" (p. 460).

The original instructions for imageability ratings were developed by Paivio et al. (1968) and were the following: "Nouns differ in their capacity to arouse mental images of things or events. Some words arouse a sensory experience, such as a mental picture or sound, very quickly and easily, whereas others may do so only with difficulty (i.e., after a long delay) or not at all. The purpose of this experiment is to rate a list of words as to the ease or difficulty with which they arouse mental images. Any word which, in your estimation, arouses a mental image (i.e., a mental picture, or sound, or other sensory experience) very quickly and easily should be given a high imagery rating: any word that arouses a mental image with difficulty or not at all should be given a low imagery rating. Think of the words "apple" or "fact." Apple would probably arouse an image relatively easily and would be rated as high imagery; fact would probably do so with difficulty and would be rated as low imagery" (p. 4).

Both sets of instruction bias toward the sense of vision. According to the concreteness instructions, something is concrete if it can be perceived through (at least one of) the senses. However, as it is has been already pointed out (Connell and Lynott, 2012, p. 461), the examples mentioned in the second part of the definition ("objects, material or persons" as well as

[1]There were also a few terms with high concreteness and low imageability ratings like e.g., "aster," "astrolabe" or "stein," but these are easy to explain: people know they denote concrete objects, but they do not know what they look like.

"chair" vs. "independence") might have biased people to rely for their ratings (also) on a different idea of concreteness which resembles more closely the everyday understanding of the word "concrete" and its dictionary definition, according to which "concrete" means material or physical and an object is concrete only if it has a material composition. Since material objects are perceived mainly or primarily through vision, people's ratings probably favored this sense over the others. Analogously to the instructions for concreteness, the instruction for imageability also evoked an idea of imageability that is primarily visual and related to the ease/difficulty with which people can form a mental picture of the referent of a word. Moreover, even though in Paivio's view "mental images" describe traces stored in memory of all kind of sensations, the term "image" recalls quite strongly the idea of a visual picture. Thus, for this aspect IMAG ratings follows criteria that overlap that of concreteness, since the instances people can more easily form a mental picture of are external, material things that they can see.

However, despite what some studies maintain (e.g., Vigliocco et al., 2009; Connell and Lynott, 2012), this is not the only relevant aspect IMAG measures. Just as CONC also measures whether/in what degree the referents of words can be experienced by senses other than sight, so IMAG measures also whether/in what degree a word arouses other kinds of sensory experience. More specifically, the request to estimate IMAG depending on whether/how much a word arouses "sensory experience" without further specifications might have lead participants to assign their ratings on the basis of the ease/difficulty with which words arouse *any kind* of sensory experience stored in memory, including internal, body-related sensations. Following Paivio et al. (1968), Paivio (1986) and Vigliocco et al. (2009), we propose that affective arousal is a kind of sensory experience, based on internal feeling rather than derived from the external senses.

A NEW HYPOTHESIS: LOOKING AT THE INCONSISTENCIES FROM AN "INTERNAL" PERSPECTIVE

This idea that word meaning might rely jointly on both internal and external sensory experience suggests that IMAG ratings might also track—at least in part—the internal and bodily-related sensory experience evoked by words. If so, IMAG diverges from CONC, and becomes a different, and more interesting measure of both the external and the internal experiential grounding of words. Since in our interpretation the imageability measure is a result of ambiguous norming instructions that lead people to assign ratings relying on their commonsense notion of sensory information, as including both internal and external information sources rather than solely external ones, we cannot assume that it is perfectly accurate. However, if we assume that people do not rely only on visual information to provide the ratings, but also spontaneously took into account their internal sensory experience and thus assigned a certain degree of IMAG to all words that aroused external and/or internal sensory experiences, then we can account for the divergence between IMAG and CONC in the case of emotion words.

If this hypothesis is correct, the class of emotion words should not be the only terminological class exhibiting a significant divergence between IMAG and CONC. In fact, all words that give rise to some kind of internal sensory experience should have an IMAG rating that is significantly higher than the CONC rating. The more a word arouses internal sensory experience, the greater should be the divergence between IMAG and CONC.

A word class that resembles emotion words insofar as it denotes body-related conditions which are experienced internally is that class denoting proprioceptive and interoceptive states. Proprioception and interoception are closely related notions: proprioception indicates our aware experience of the position of our body (see e.g., Berthoz, 2000); while interoception describe people's general conscious experience of their bodily states or of specific conditions of parts of their body (Craig, 2003, 2009, 2010). Words describing typical proprioceptive states and interoceptive states are e.g., balance, relaxation, movement, tremor, sit, rest, jump, run, walk etc. on the one hand and on the other ache, sick, hunger, thirsty, warmth, itch, pain, cold, etc.

Emotion, proprioceptive, and interoceptive words might however not be the only ones relying on internal, bodily-related sensory experience. In fact, some recent studies carried out in the field of so called embodied cognition suggest that abstract words are also grounded in internal states, especially affective and mental states (see e.g., Barsalou and Wiemer-Hastings, 2005; Kousta et al., 2009, 2011; Vigliocco et al., 2009; for a review of older studies see e.g., Barsalou, 1999, p. 599). In particular, the studies by Wiemer-Hastings et al. show that abstract words tend to have more introspective and affective associations than concrete words (Wiemer-Hastings and Xu, 2005; Vigliocco et al., 2009; Kousta et al., 2011). As these studies suggest, abstract concepts clearly cannot rely only on affective information, their representation must also be based on linguistic information, and the exact proportion of affective and linguistic information will vary depending on the word (see e.g., Vigliocco et al., 2009; Kousta et al., 2011). However, if we admit that abstract words do indeed also rely at least minimally on internal sensory experience and hypothesize that IMAG ratings measure whether a word arouses internal sensory experience, then in the case of abstract words the correlation between CONC and IMAG should be significantly smaller than in the case of concrete words because IMAG ratings should be relatively higher than CONC ratings.

Some results in line with this prediction were already reported by Altarriba et al. (1999) and by Wiemer-Hastings et al. (2001); however a more accurate analysis is needed. According to our hypothesis, correlation patterns should differ when calculated separately for decreasing CONC ratings: the more abstract a word is, the weaker the correlation between CONC and IMAG should become. In addition, since the proportion of affective and linguistic information abstract words rely on varies depending on the kind of words we are considering, we expect that highly theoretical words with a technical meaning that have only a limited everyday use and strictly depend on their linguistic definition (e.g., adverb, literal, plenipotentiary, causality, regulation, abduction, deduction, axiom, factor, fallacy, function, suffrage etc.) will have relatively low IMAG ratings with respect to CONC ratings. More specifically, since their proportion of linguistic information is particularly high in comparison to sensory information, if our hypothesis about IMAG is correct, the difference between the IMAG ratings and the CONC ratings for this class of words

should be either smaller than, or comparable to, that of other word groups.

TESTING THE HYPOTHESIS

To prove our hypothesis, we analyzed the CONC and the IMAG ratings included in the MRC database, which is an important source for these measures in psycholinguistic studies and constitutes the only database available in which CONC and IMAG were collected simultaneously by the same studies. These not only rely on exactly the same instruction we discussed previously, but were also driven by the intent of understanding the relationship between IMAG and CONC.

Later collection of IMAG and CONC ratings available in English are not directly relevant with respect to our hypothesis for a number of reasons. First of all, recent collections of IMAG ratings do not also include CONC ratings (Bird et al., 2001; Cortese and Fugett, 2004; Stadthagen-Gonzalez and Davis, 2006; Schock et al., 2012). Since the clue to understand the theoretical peculiarities of the construct of imageability resides in the anomalies with respect to concreteness, it is by considering the imageability ratings in relation to the concreteness ratings—i.e., by comparing them—that a new insight into the construct of imageability can be achieved. Secondly, some collections rely on instructions that differ at least in some respect from the one we discussed: this is the case for the recently published database of CONC (Brysbaert et al., 2013) as well as for the collection of IMAG ratings carried out by Bird et al. (2001)[2]. Thirdly, some of these collections are very specific in scope and consider only monosyllabic and disyllabic words (Cortese and Fugett, 2004; Schock et al., 2012). Finally, Stadthagen-Gonzalez's and Davis' database is obtained merging their data with Gilhooly and Logie's (1980) collection, which is already included in the MRC database.

The MRC psycholinguistic database includes 9240 words possessing an IMAG rating and 8228 words possessing a CONC rating. Both are derived from merging three sets of norms: the Colorado Norms (Toglia and Battig, 1978), the Pavio Norms (unpublished, these are an expansion of the norms of Paivio et al., 1968), and the Gilhooly-Logie norms (Gilhooly and Logie, 1980). A large part of the data from Toglia and Battig (1978) was validated by Cortese and Fugett (2004). The values are in the range 100–700. Words are partitioned in ten syntactic categories: nouns, adjectives, verbs, adverbs, conjunctions, pronouns, interjections, past participles, other.

SELECTION OF STIMULI

For our analysis we considered only words that have both an IMAG and a CONC rating and we excluded conjunctions, pronouns, interjections, and the class labeled "other." Repetitions were also excluded. This leaves 4260 words.

Across all words, mean IMAG and CONC ratings are 456.4 and 438.7 respectively. The correlation between IMAG and CONC is significant ($r = 0.835$, $p < 0.001$), which demonstrates—as has been previously observed—that the two constructs are tightly interconnected. Interestingly, if two groups of words are construed as a function of CONC (low vs. high CONC ratings, 2130 words in each group; mean CONC and IMAG ratings for the low CONC group: 331.6 and 376.6, respectively; mean CONC and IMAG ratings for the high CONC group: 545.8 and 536.3, respectively), the correlation between IMAG and CONC for the low CONC group ($r = 0.550$, $p < 0.001$) is significantly smaller than the correlation between IMAG and CONC for the high CONC group ($r = 0.661$, $p < 0.001$), $z = 5.7$, $p < 0.001$[3]. This is compatible with the view that IMAG ratings are less dependent upon CONC ratings for the abstract (i.e., low concrete) words with respect to the concrete words. Since, as specified in section Instruction-Bound Ratings?, abstract words generally rely on more introspective information than concrete words, these different correlation patterns suggest that IMAG does not entirely depend on CONC, but it is also a measure of something else, and specifically of the ease/difficulty with which a word evokes internal sensory experience of any kind (be it e.g., emotional, proprioceptive or interoceptive).

To test the hypothesis that IMAG ratings depend on the ease/difficulty with which a word arouses both external and/or internal sensory experience, and that a discrepancy between CONC and IMAG may be diagnostic of the relative contribution of the two kinds of sensory information, we selected three groups of words: (i) 36 emotional words (whose anomalous behavior as for their IMAG and CONC ratings has already been singled out by other studies—on this point see section Imageability in the MRC Psycholinguistic Database), (ii) 56 proprioceptive or interoceptive words (which we call globally *X-ceptive* to indicate that the same considerations we develop for proprioception and interoception should apply for any kind of states based on an internal perception), and (iii) 110 theoretical terms (i.e., abstract technical terms whose meaning is not grounded on internal states, but depends rather on a linguistic definition given in the framework of a theory). In addition, we construed ten control groups of 100 randomly selected words to compare with (i), (ii), and (iii). Selection of the words to serve in the control groups was accomplished through a computerized algorithm, with the only restriction that none of the words in the emotional, X-ceptive or theoretical group could be selected to serve in the control groups.

(i) The class of emotion words combines two kinds of words: those strictly denoting emotions and those denoting what are more correctly called moods (or background feelings—e.g., depression, anxiety, wellness, distress, etc.). In order to individuate a particularly salient and unambiguous set of terms, our selection from the MRC database was based on the emotions/moods described by a number of studies (which sometimes consider a mixture of the two). As for emotions, we included only emotions considered as basic (Tomkins, 1962, 1963; Ekman et al., 1969; Plutchik, 1980; Ekman, 1999; Reizenzein, 2009; Kassam et al.,

[2]The new instructions for concreteness are particularly problematic since they ask people to evaluate actions as something concrete equating them from a semantic point of view with objects. A study specifically devoted to the role of instructions in collecting useful psycholinguistic data on concreteness is in preparation.

[3]Correlation were compared according to the following procedure. First *r* values have been converted into *z* values (Fisher *r* to *z* transformation), then the results have been compared taking into consideration the sample size (Cohen and Cohen, 1983).

2013). While the words denoting basic emotions are all strictly derived from the mentioned studies, the list of the words denoting moods is more freely composed starting from the examples and the definitions given in various studies (Ekman, 1994; Damasio, 1999; Prinz, 2004). Emotion words strongly rely on internal affective experience; thus, if IMAG ratings measure how easily a word evokes not only external but also internal sensory information, IMAG ratings for this class of words should be significantly higher than CONC ratings compared to the other groups of words.

(ii) Words denoting proprioceptive and interoceptive (X-ceptive) states were selected from the MRC database starting from the examples considered in the studies of Berthoz (2000) and Craig (2003, 2009, 2010). Since proprioceptive and interoceptive (X-ceptive) states are analogous to emotions due to the fact that they are based on internal sensory experience, we expect words denoting these states to behave like emotion words and exhibit IMAG ratings significantly higher than CONC ratings with respect to other groups of words.

iii) For the selection of theoretical words we could not rely on previous studies, even though the definition of a class of theoretical terms as opposed to a class of observational terms was already introduced by Paivio (Paivio, 1986, p. 10 Clark and Paivio, 1989). However, while Paivio interpreted theoretical terms simply as abstract terms, we consider theoretical terms as an autonomous subclass of abstract words. In our account, theoretical terms are technical words with a definitional structure whose meaning is fixed in the framework of a theory. We identified this group of words one by one in the database according to this criterion: the chosen terms belong to the technical jargon of a discipline and therefore strictly depend on a specific linguistic definition. Thus, we avoided terms that denote anything that can be perceived through the senses. An example is the mathematical term "axiom," i.e., a statement or formula on which an abstractly defined structure is based. Other than from mathematics, terms come from physics (e.g., "causality"), linguistics (e.g., "conjugation"), politics and law (e.g., "legislation"), logic (e.g., "deduction"), and science in general (e.g., "theory"). Since this class should rely only very weakly on internal sensory experience, will have relatively low IMAG ratings with respect to CONC ratings. Specifically, we expect that the difference between the IMAG ratings and the CONC ratings for this class of words will be either smaller than, or comparable to, that of other word groups (control groups as well as emotion and X-ception words).

PROCEDURE

We compared the differences between ratings of IMAG and CONC of these three groups against the differences between the ratings of IMAG and CONC of ten control groups including 100 randomly selected words (basically, a bootstrap). The idea here is that the mean differences between IMAG and CONC of the control groups—being composed of randomly selected words—reflect the mean differences between IMAG and CONC of the population they derived from. Therefore, if one (or more) of the three experimental groups consistently and significantly differs from the control group(s), then we can conclude that that experimental group differs from the population on the tested dimensions. The comparisons were made using an ANOVA with Group [experimental (X-ception, emotion, or technical) vs. control (each of the 10 control groups)] as a between-items factor. In addition, each of the experimental groups was compared with the other two experimental groups.

RESULTS

The results are reported in **Table 1**. The first 10 rows refer to the comparison of each experimental group of words with one of the control groups. The last two rows refer to the comparison among the experimental groups of words.

As expected, the differences between IMAG and CONC for the X-ception words are significantly higher than those of the control groups. Also, and in line with previous evidence, the differences between IMAG and CONC for the emotion words are significantly higher than those of the control groups. In addition, and congruently with the theory at the basis of our hypothesis, the differences between IMAG and CONC for the theoretical/technical words are either smaller than, or comparable to, those of the control groups. Unexpectedly, the differences between IMAG and CONC of the X-ception words are significantly smaller than those of the emotion words: This will be dealt with in the General Discussion.

GENERAL DISCUSSION

According to the hypothesis we put forward, IMAG is a construct based on two factors. On the one hand, IMAG depends on CONC, since it measures the ease/difficulty with which a word evokes external (mainly visual) sensory experience related to the objects it denotes. On the other hand, IMAG is partially independent of CONC and measures the ease/difficulty with which a word evokes internal sensory experience. To test this hypothesis we used the IMAG and CONC ratings included in the MRC database. Since we assume that IMAG is always linked to CONC and that IMAG ratings will therefore always reflect to a certain extent CONC ratings, we are not interested in analyzing IMAG and CONC ratings *per se*, but we focus on the difference between IMAG and CONC ratings which reveals a value of IMAG independent from CONC.

Our analysis started from some basic assumptions regarding what type of information different classes of words rely on. Indeed, words might be grounded on both external or internal sensory information: (a) while concrete words rely primarily on external sensory information, in general abstract words are mainly based, to a larger or smaller extent, on internal sensory information and on linguistic information. (b) Among the abstract terms that surely rely for a large part on internal sensory information there are words denoting emotions as well as proprioceptive and interoceptive states. (c) On the contrary, theoretical words denoting technical notions will probably be mainly linguistic constructs and be based only to a very small extent on sensory information of any kind.

(a) Moving from this premise, we examined first of all whether in general the correlation between CONC and IMAG varies for concrete and abstract words. While in the case of concrete words, IMAG ratings should be just a function of CONC ratings, in the case of abstract words which rely to a certain degree on internal information the correlation between CONC and IMAG should be significantly smaller. Our analysis on two groups of words constructed as a function of CONC confirmed this hypothesis.

Table 1 | Results of the ANOVAs.

			X-ceptions $N = 56$ $M = 72.4$ $CI = 62, 83$	Emotion $N = 36$ $M = 119.0$ $CI = 95, 143$	Theoretical $N = 110$ $M = -1.3$ $CI = -12, 9$
Control groups	1	*F*	45.7	75.8	3.0
	N = 100	dof	1, 154	1, 134	1, 208
	M = 12.4	MSE	2829	3963	3280
	CI = 0.6, 24	*P*	<0.001	<0.001	=0.083
	2	*F*	51.8	77.4	0.1
	N = 100	dof	1, 154	1, 134	1, 208
	M = 1.4	MSE	3494	4728	3773
	CI = −12, 15	*P*	<0.001	<0.001	=0.745
	3	*F*	23.6	48.9	8.0
	N = 100	dof	1, 154	1, 134	1, 208
	M = 23.1	MSE	3704	4969	3929
	CI = 9, 37	*P*	<0.001	<0.001	<0.01
	4	*F*	30.8	58.0	5.6
	N = 100	dof	1, 154	1, 134	1, 208
	M = 18.5	MSE	3396	4615	3700
	CI = 5, 32	*P*	<0.001	<0.001	<0.05
	5	*F*	25.2	54.5	12.2
	N = 100	dof	1, 154	1, 134	1, 208
	M = 26.8	MSE	2973	4128	3387
	CI = 15, 39	*P*	<0.001	<0.001	<0.005
	6	*F*	37.3	63.3	2.1
	N = 100	dof	1, 154	1, 134	1, 208
	M = 11.1	MSE	3615	4866	3862
	CI = −3, 25	*P*	<0.001	<0.001	=0.147
	7	*F*	46.2	74.1	1.4
	N = 100	dof	1, 154	1, 134	1, 208
	M = 8.54	MSE	3171	4356	3533
	CI = −4, 21	*p*	<0.001	<0.001	=0.231
	8	*F*	47.3	74.5	0.9
	N = 100	dof	1, 154	1, 134	1, 208
	M = 6.7	MSE	3276	4477	3611
	CI = −6, 20	*P*	<0.001	<0.001	=0.332
	9	*F*	30.8	59.0	6.7
	N = 100	dof	1, 154	1, 134	1, 208
	M = 19.9	MSE	3209	4400	3562
	CI = 7, 32	*P*	<0.001	<0.001	<0.05
	10	*F*	48.9	78.5	2.0
	N = 100	dof	1, 154	1, 134	1, 208
	M = 0.85	MSE	2878	4020	3317
	CI = −2, 22	*P*	<0.001	<0.001	=0.161
	X-ceptions	*F*		16.2	79.9
	N = 56	dof		1, 90	1, 164
	M = 72.4	MSE		2926	2528
	CI = 62, 83	*P*		<0.001	<0.001
	Emotion	*F*			110.9
	N = 36	dof			1, 144
	M = 119.0	MSE			3542
	CI = 95, 143	*P*			<0.001

N, size of the sample; M, mean differences between imageability and concreteness ratings (ratings from the MRC psycholinguistic database); CI, Confidence Intervals (95%); Once corrected for multiple comparisons, separately for each conceptual category, $\alpha = 0.005$.

(b) Secondly, we selected two groups of words from the database, one denoting emotions and the other proprioceptive/interoceptive (x-ceptive) states. Since these classes of words rely to a large degree on internal information, we expected that the difference between IMAG and CONC for both classes would be significantly higher than that for the control groups. Our statistical analysis supports this hypothesis.

An unexpected finding here is that the difference between IMAG and CONC for the X-ception words is significantly smaller than that for the emotion words. There are at least two possible explanations for this result.

First, one could speculate that it is due to the fact that emotions rely on internal bodily information in a twofold manner. On the one hand, an emotional state is revealed by a specific affective arousal and internal feelings. On the other hand, emotions bring about specific bodily reactions and above all specific facial expressions which are an essential part of emotions (see e.g., Ekman, 1984) and are recorded through interoception giving additional bodily information on the state. Thus, one could hypothesize that the higher difference between IMAG and CONC for emotion words compared with x-ceptive words is due to this double binding between emotions and internal sensory experience: in this case, mean IMAG ratings for emotion words should be higher than those for X-ception words, whereas mean CONC ratings for the two classes of words should be similar. A second possible explanation is that the sensory information corresponding to proprioceptive and interoceptive states (i.e., internal world perception) is "qualitatively comparable" to (external world-) perception and is therefore interpreted as more concrete than the affective arousal/feelings corresponding to emotions. That is, people could consider words like "ache," "hunger," "cold," "hot," "motion," "itch" as denoting more tangible and specific (i.e., concrete) states than words like "happiness," "sadness," "excitement," "humiliation," "jealousy" etc. As a consequence, in this case CONC ratings of X-ceptive words should be higher than those of emotional words, whereas mean IMAG ratings for the two classes of words should be similar.

To distinguish between these two hypotheses, we performed two ANOVAs. In one analysis, we compared the CONC ratings of X-ception and emotion words. This analysis showed that the concreteness ratings of X-ception words was significantly higher than the concreteness ratings of emotion words (means: 391 vs. 314, respectively; $[F_{(1,\ 90)} = 39.4, \text{MSE} = 3334, p < 0.001]$. In the second analysis we compared the IMAG ratings of X-ception and emotion words. This second analysis showed that the imageability ratings of X-ception words was significantly higher than the imageability ratings of emotion words [464 vs. 433, respectively, $F_{(1,\ 90)} = 7.2, \text{MSE} = 2904, p < 0.01$]. Unfortunately, these analyses do not allow us to conclusively decide in favor of either of the two hypotheses put forward above, since both IMAG and CONC ratings are lower for emotion words with respect to X-ception words. It is worth noting that the difference between the mean CONC ratings of the two classes of words is larger than the difference between the mean IMAG ratings, and this, if anything, provides (weak) support for the second of our hypotheses.

(c) Finally, we selected from the database a group of theoretical/technical words which should only weakly rely on

internal information and therefore have relatively low IMAG ratings with respect to CONC ratings. Congruently with this hypothesis, the differences between IMAG and CONC for the theoretical/technical words turned out to be either smaller (4 out of 10 comparisons are significant if $\alpha = 0.05$; 1 out of 10 comparisons are significant once α is corrected for multiple comparisons[4]; c.f. **Table 1**) then or comparable to those of the control groups.

Taken together, these results show that IMAG is not simply an alternative way to measure concreteness, but, instead, that IMAG provides specific information and depends in part on the strength with which words evoke body-internal sensations. We think that this result is extremely useful, among other things, for better understanding how to use IMAG and CONC ratings for experimental research.

One of the main applications of these ratings is in studies that analyze the processing advantages of some classes of words over others; most famously, the processing advantages of concrete vs. abstract words (the so-called concreteness effect). In this case, our results suggest not only that the two ratings should not be used interchangeably, but they also indicate that—in addition to a concreteness effect—it might be possible to identify an effect specifically related to imageability (and more precisely to "the side" of imageability that does not depend on concreteness), which measures the ease/difficulty with which words evoke some kind of internal sensory information. In this case, processing advantages should be observed for both emotion and X-ceptive words.

CONCLUDING REMARKS

In this paper we analyzed the IMAG and CONC ratings included in the MRC Psycholinguistic Database. We started by presenting the results of some previous studies showing that—even though there is a strong correlation between measures of IMAG and CONC—some words with low CONC ratings (i.e., abstract words) exhibit a contrasting pattern and have relatively high IMAG ratings. Also on the basis of an analysis of the instructions given to subjects during the collection of the ratings, we hypothesized that IMAG is only in part connected to CONC; while to a certain extent is independent from it and measures something different: i.e., the ease/difficulty with which a word arouses any kind of sensory experience including internal, body-related sensations. In order to validate this position, we carried out several analyses of the IMAG and CONC ratings in the database, individuating different groups of words and considering for each the difference between IMAG and CONC. All the results are congruent with the initial hypotheses.

These results show that IMAG ratings depend at least on two factors: i.e., on the one hand, on whether a word denotes concrete external objects (and for this aspect IMAG directly relies on CONC) and on the other, on whether a word is grounded in any kind of internal, body-related sensations. As we showed, this is the case for words denoting e.g., emotions as well as proprioceptive and interoceptive states.

This conclusion serves not only to reaffirm at least to some degree the reliability of the IMAG measure, in spite of the well-known inconsistencies that characterize it, which we interpreted from an entirely new perspective, but it also has relevant consequences at least with respect to two different points. On the one hand, it challenges the widely shared idea that CONC and IMAG are interchangeable scales measuring one and the same thing. On the other hand, our analysis helps to clarify the IMAG construct and to specify what it exactly measures. This has direct implication e.g. for the debate on the relationship between abstract and concrete. According to Vigliocco, Kousta, and collaborators (Vigliocco et al., 2009, 2013; Kousta et al., 2011) the dichotomy between abstract and concrete words is only apparent, as word meanings rely in different proportions on perception, internal information, and linguistic information: thus, generally people call "concrete" the words with an higher proportion of perceptual information, while they call "abstract" the words relying primarily on internal and linguistic information. Since imageability ratings are a joint measure of the link of words with both external and internal sensory experience, we suggest that the extent of a word's connection with internal sensory information can be obtained by subtracting its concreteness rating from its imageability rating. If the interpretation of imageability we put forward is correct, then the intersection between imageability and concreteness can indeed give us a tangible (even though not completely accurate) measure of the internal sensory information aroused by a word.

Our analysis shows also that the imageability and even concreteness measures are complex and problematic constructs, whose ratings undergo biases that cannot be completely controlled. New collections of ratings on the basis of less ambiguous instructions are required in order to have more precise measures to use for experimentation, i.e., to show among other things whether an abstract word mainly rely on linguistic information and is therefore theoretical or whether also an abstract word is strongly grounded in internal information.

[4]Note that correcting for multiple comparisons is not necessary in this context. We reported the corrected values to show that even with a more conservative approach, the semantic of the pattern of results does not change.

REFERENCES

Altarriba, J., and Bauer, L. M. (2004). The distinctiveness of emotion concepts: a comparison between emotion, abstract, and concrete words. *Am. J. Psychol.* 117, 389–410. doi: 10.2307/4149007

Altarriba, J., Bauer, L. M., and Benvenuto, C. (1999). Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words. *Behav. Res. Methods* 31, 578–602. doi: 10.3758/BF03200738

Barsalou, L. W. (1999). Perceptual symbol systems. *Behav. Brain Sci.* 22, 577–660.

Barsalou, L. W., and Wiemer-Hastings, K. (2005). "Situating abstract concepts," in *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thought,* eds D. Pecher and R. Zwaan (New York, NY: Cambridge University Press), 129–163.

Benjafield, J., and Muckenheim, R. (1989). Dates of entry and measures of imagery, concreteness, goodness, and familiarity for 1046 words sampled from the Oxford English Dictionary. *Behav. Res. Methods Instrum. Comput.* 21, 31–52.

Berthoz, A. (2000). *The Brain's Sense of Movement*. Cambridge: Harvard University Press.

Bird, H., Franklin, S., and Howard, D. (2001). Age of acquisition and imageability ratings for a large set of words, including verbs, and function words. *Behav. Res. Methods Instrum. Comput.* 33, 73–79. doi: 10.3758/BF03195349

Brysbaert, M., Warriner, A. B., and Kuperman, V. (2013). Concreteness ratings for 40 thousand generally known English word lemmas. *Behav. Res. Methods.* doi: 10.3758/s13428-013-1403-5. [Epub ahead of print].

Christian, J., Bickley, W., Tarka, M., and Clayton, K. (1978). Measures of free recall of 900 English nouns: correlations with imagery, concreteness, meaningfulness, and frequency. *Mem. Cogn.* 6, 379–390. doi: 10.3758/BF03197470

Clark, J. M., and Paivio, A. (1989). Observational and theoretical terms in psychology. *Am. Psychol.* 44, 500–512. doi: 10.1037/0003-066X.44.3.500

Cohen, J., and Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 2nd Edn.* Hillsdale, NJ: Erlbaum.

Coltheart, M. (1981). The MRC psycholinguistic database. *Q. J. Exp. Psychol. Sec. A Hum. Exp. Psychol.* 33, 497–505. doi: 10.1080/14640748108400805

Connell, L., and Lynott, D. (2012). Strength of perceptual experience predicts word processing performance better than concreteness or imageability. *Cognition* 125, 452–465. doi: 10.1016/j.cognition.2012.07.010

Cortese, M. J., and Fugett, A. (2004). Imageability ratings for 3,000 monosyllabic words. *Behav. Res. Methods Instrum. Comput.* 36, 384–387. doi: 10.3758/BF03195585

Craig, A. D. (2003). Interoception: the sense of the physiological condition of the body. *Curr. Opin. Neurobiol.* 13, 500–505. doi: 10.1016/S0959-4388(03)00090-4

Craig, A. D. (2009). How do you feel – now? The anterior insula and human awareness. *Nat. Rev. Neurosci.* 10, 59–70. doi: 10.1038/nrn2555

Craig, A. D. (2010). The Sentient Self. *Brain Struct. Funct.* 214, 563–577. doi: 10.1007/s00429-010-0248-y

Damasio, A. (1999). *The Feeling of What Happens. Body and Emotion in Making of Consciousness*. New York, NY: Mariner Books.

Dellantonio, S., Job, R., and Mulatti, C. (2014). Imageability: now you see it again (albeit in a different form). *Front. Psychol.* 5:279. doi: 10.3389/fpsyg.2014.00279

Ekman, P. (1984). "Expression and the nature of emotion," in *Approaches to Emotion,* eds K. Scherer and P. Ekman (Hillsdale, NJ: Lawrence Erlbaum), 319–343.

Ekman, P. (1994). "Moods, emotions and traits," in *The Nature of Emotion: Fundamental Questions*, eds P. Ekman and R. J. Davidson (Oxford: Oxford University Press), 56–58.

Ekman, P. (1999). "Basic emotions," in *Handbook of Cognition and Emotion*, *Chapter 3,* eds T. Dalgleish and M. Power (Sussex: John Wiley & Sons), 45–60.

Ekman, P., Sorenson, E. R., and Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science* 164, 86–88. doi: 10.1126/science.164.3875.86

Friendly, M., Franklin, P. E., Hoffman, D., and Rubin, D. C. (1982). The toronto word pool: norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behav. Res. Methods Instrum.* 14, 375–399. doi: 10.3758/BF03203275

Gilhooly, K. J., and Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behav. Res. Methods Instrum.* 12, 395–427. doi: 10.3758/BF03201693

Kassam, K. S., Markey, A. R., Cherkassy, V. L., Loewenstein, G., and Just, M. A. (2013). Identifying emotions on the basis of neural activation. *PLoS ONE* 8:e66032. doi: 10.1371/journal.pone.0066032

Kousta, S. T., Vigliocco, G., Vinson, D. P., Andrew, M., and Del Campo, E. (2011). The representation of abstract words: why emotion matters. *J. Exp. Psychol. Gen.* 140, 14–34. doi: 10.1037/a0021446

Kousta, S. T., Vinson, D. P., and Vigliocco, G. (2009). Emotion words, regardless of polarity, have a processing advantage over neutral words. *Cognition* 112, 473–481. doi: 10.1016/j.cognition.2009.06.007

Paivio, A. (1971). *Imagery and Verbal Processes*. New York, NY: Holt, Rinehart & Winston.

Paivio, A. (1986). *Mental Representations. A Dual Coding Approach*. Oxford: Oxford University Press.

Paivio, A. (2007). *Mind and its Evolution: A Dual Coding Theoretical Approach*. Mahwah, NJ: Erlbaum.

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

Plutchik, R. (1980). "A general psychoevolutionary theory of emotion," in *Emotion: Theory, Research, and Experience: Theories of Emotion,* Vol. 1, eds R. Plutchik and H. Kellerman (New York, NY: Academic), 3–33.

Prinz, J. J. (2004). *Gut Reactions. A Perceptual Theory of Emotion*. Oxford: Oxford University Press.

Reilly, J., and Kean, J. (2007). Formal distinctiveness of high and low imageability nouns: analyses and theoretical implications. *Cogn. Sci.* 31, 1–12. doi: 10.1080/03640210709336988

Reizenzein, R. (2009). Emotional experience in the computational belief-desire theory of emotion. *Emot. Rev.* 1, 214–222. doi: 10.1177/1754073909103589

Rubin, D. C. (1980). Fifty-one properties of 125 words: a unit analysis of verbal behavior. *J. Verb. Learn. Verb. Behav.* 19, 736–755. doi: 10.1016/S0022-5371(80)90415-6

Rubin, D. C., and Friendly, M. (1986). Predicting which words get recalled: measures of free recall, availability, goodness, emotionality, and pronunciability for 925 nouns. *Mem. Cognit.* 14, 79–94. doi: 10.3758/BF03209231

Schock, J., Cortese, M. J., and Khanna, M. M. (2012). Imageability estimates for 3,000 disyllabic words. *Behav. Res.* 44, 374–379. doi: 10.3758/s13428-011-0162-0

Schwanenflugel, P. J., Harnishfeger, K. K., and Stowe, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *J. Mem. Lang.* 27, 499–520. doi: 10.1016/0749-596X(88)90022-8

Spreen, O., and Schulz, R. W. (1966). Parameters of abstraction, meaningfulness, and pronunciability for 329 nouns. *J. Verb. Learn. Verb. Behav.* 5, 459–468. doi: 10.1016/S0022-5371(66)80061-0

Stadthagen-Gonzalez, H., and Davis, C. J. (2006). The bristol norms for age of acquisition, imageability, and familiarity. *Behav. Res. Methods* 38, 598–605. doi: 10.3758/BF03193891

Toglia, M. P., and Battig, W. F. (1978). *Handbook of Semantic Word Norms*. Hillsdale, NJ: Erlbaum.

Tomkins, S. S. (1962). *Affect, Imagery, Consciousness: Vol. I: The Positive Affects*. New York, NY: Springer.

Tomkins, S. S. (1963). *Affect, Imagery, Consciousness: Vol. II: The Negative Affects*. New York, NY: Springer.

Vigliocco, G., Kousta, S., Vinson, D., Andrews, M., and Del Campo, E. (2013). The representation of abstract words: What matters? Reply to Paivio's (2013) comment on Kousta et al. (2011). *J. Exp. Psychol. Gen.* 142, 288–291. doi: 10.1037/a0028749

Vigliocco, G., Meteyard, L., Andrews, M., and Kousta, S. (2009). Toward a theory of semantic representation. *Lang. Cogn.* 1, 219–247. doi: 10.1515/LANGCOG.2009.011

Westbury, C. F., Shaoul, C., Hollis, G., Smithson, L., Briesemeister, B. B., Hofmann, M. J., et al. (2013). Now you see it, now you don't: on emotion, context, & the algorithmic prediction of human imageability judgments. *Front. Psychol.* 4:991. doi: 10.3389/fpsyg.2013.00991

Wiemer-Hastings, K., Krug, J., and Xu, X. (2001). "Imagery, context availability, contextual constraint and abstractness," in *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society,* eds J. D. Moore and K. Stenning (Edinburgh), 1106–1111.

Wiemer-Hastings, K., and Xu, X. (2005). Content differences for abstract and concrete concepts. *Cogn. Sci.* 29, 719–773. doi: 10.1207/s15516709cog0000_33

Wilson, M. D. (1988). The MRC Psycholinguistic database: machine readable dictionary, Version 2. *Behav. Res. Methods Instrum. Comput.* 20, 6–11. doi: 10.3758/BF03202594

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.

Received: 02 March 2014; accepted: 19 June 2014; published online: 15 July 2014.
Citation: Dellantonio S, Mulatti C, Pastore L and Job R (2014) Measuring inconsistencies can lead you forward: Imageability and the x-ception theory. Front. Psychol. ***5****:708. doi: 10.3389/fpsyg.2014.00708*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

METHODS ARTICLE
published: 31 March 2014
doi: 10.3389/fpsyg.2014.00267

Hierarchical clustering analysis of reading aloud data: a new technique for evaluating the performance of computational models

Serje Robidoux* and Stephen C. Pritchard

ARC Centre of Excellence in Cognition and its Disorders, Department of Cognitive Science, Macquarie University, Sydney, NSW, Australia

Edited by:
Davide Crepaldi, University of Milano-Bicocca, Italy

Reviewed by:
Claudio Mulatti, University of Padova, Italy
Michael E. J. Masson, University of Victoria, Canada

****Correspondence:***
Serje Robidoux, ARC Centre of Excellence in Cognition and its Disorders, Department of Cognitive Science, Australian Hearing Hub, Level 3, 16 University Avenue, Macquarie University, NSW 2109, Australia
e-mail: serje.robidoux@mq.edu.au

DRC (Coltheart et al., 2001) and CDP++ (Perry et al., 2010) are two of the most successful models of reading aloud. These models differ primarily in how their sublexical systems convert letter strings into phonological codes. DRC adopts a set of grapheme-to-phoneme conversion rules (GPCs) while CDP++ uses a simple trained network that has been exposed to a combination of rules and the spellings and pronunciations of known words. Thus far the debate between fixed rules and learned associations has largely emphasized reaction time experiments, error rates in dyslexias, and item-level variance from large-scale databases. Recently, Pritchard et al. (2012) examined the models' non-word reading in a new way. They compared responses produced by the models to those produced by 45 skilled readers. Their item-by-item analysis is informative, but leaves open some questions that can be addressed with a different technique. Using hierarchical clustering techniques, we first examined the subject data to identify if there are classes of subjects that are similar to each other in their overall response profiles. We found that there are indeed two groups of subject that differ in their pronunciations for certain consonant clusters. We also tested the possibility that CDP++ is modeling one set of subjects well, while DRC is modeling a different set of subjects. We found that CDP++ does not fit any human reader's response pattern very well, while DRC fits the human readers as well as or better than any other reader.

Keywords: computational modeling, reading aloud, hierarchical clustering, non-word reading

Reading aloud involves converting printed character strings into phonological codes. In the case of words, one can rely on memory structures to provide the appropriate pronunciation. However, where novel letter strings are concerned the reader must perform the translation in some other way. One question under considerable debate is whether readers adopt a set of strictly applied rules for this conversion, or if there is a more subtle set of associative relationships between letter patterns and pronunciation at play. The role of grapheme-phoneme conversion rules (or GPCs) and trained, neural networks that learn implicit associations is at the heart of the debate between two of the most broadly successful computational models of reading aloud currently available. When converting printed words into phonology, both the Dual-Route-Cascaded Model (DRC; Coltheart et al., 2001) and the Connectionist Dual-Process models of reading aloud (CDP+/++; Perry et al., 2007, 2010) rely principally on nearly identical lexical systems that store the appropriate information. When it comes to pronounceable non-word letter strings, however, DRC assumes reading is accomplished through the use of GPCs, whilst CDP+ and CDP++ rely instead on a simple neural network that has learned to associate graphemes with phonemes through exposure to a combination of real words and rules.

In debating the relative merits of the two approaches, researchers have relied extensively on experimental results that used reaction times and error rates as the principal variables of interest. On those metrics, CDP++ enjoys an advantage over DRC: it is able to simulate consistency effects, and is able to account for more of the variance in human response times when assessed against large-scale database studies such as the English Lexicon Project (Perry et al., 2007).

While these modal metrics of human behavior are important, they ignore a separate question that is particularly relevant to the debate between strong GPCs such as those in DRC and associative learning algorithms such as the one implemented in CDP++: do the pronunciations produced by the models in response to novel stimuli match those of human readers? In other words, when presented with an item like "PHLOMB," DRC produces the response /flɔm/,[1] (as in "bomb") while CDP++ responds /fləʊm/ (as in "comb"). Little research has thus far compared the model responses to those produced by subjects.

Pritchard et al. (2012) examined just this question. They submitted 1475 non-word letter strings made up of onsets and bodies that exist in English, and legal bigrams, to DRC and CDP+/++ and identified 412 that differentiated between the two models. 45 human readers then read these 412 items aloud and their responses were coded for phonology. Comparing the human responses to those of the models, they found that, while both models had some difficulties in matching the empirical data, DRC

[1]The notation used here is based on the International Phonetic Alphabet.

outperformed CDP+ and CDP++. For these 412 items, DRC was more likely to produce the response most common among the subjects (the modal response), and less likely to produce a unique response that no human reader produced.[2]

AN UNANSWERED QUESTION

Pritchard et al.'s (2012) item-by-item analysis clearly favors the view that DRC captures "typical" human non-word reading better than CDP+ or CDP++, but it's difficult to know what "typical" means here. Subjects vary considerably in the kinds of responses they produce. Whereas some non-words produced 100% agreement among the subjects, other non-words resulted in up to 24 different responses. This difficulty led Rastle and Coltheart (1999) to define the DRC's goal as producing the modal response for all items:

> "All we seek to achieve is that for all non-words, the DRC model's pronunciation is the one that the majority of readers assign." (p. 484)

However, it is evident even from the Pritchard et al. (2012) data that this is not always possible: twelve items produced more than one possible modal response (e.g., SLYS was pronounced as "sleece," "slice," and "sleeze" by 12 readers each). In other cases, though there was one true mode in the sample, there was often a very near-modal alternative response (e.g., CESH is pronounced as "sesh" by 19 subjects, and "kesh" by 20 subjects). For such items, choosing the target response according Rastle and Coltheart's (1999) goal is not as unambiguous as it might at first seem.

An alternative (and probably complementary) approach to evaluating the model success is to compare subject response *profiles* against each other and against the models to determine whether there are different groups of subjects with similar response profiles, and whether the models perform better at fitting some of these groups over others. Looking at overall profiles rather than item-by-item analyses allows us to ask two questions: first, are there clusters of subjects that tend to respond similarly in a way that is not readily detected by the item-by-item approach. Second, are there some subjects who seem to match the DRC's GPC-driven responses while others tend to use the more fluid associations learned by CDP+/++? Answering this question requires a way to simultaneously compare all subjects and the models on their overall response profiles, and not on an item-by-item basis. Here we discuss one approach to this problem.

HIERARCHICAL CLUSTERING

Hierarchical clustering techniques are designed to do just this. Conceptually, hierarchical clustering[3] is a simple algorithm:

1. For each possible pair of subjects, produce an index of how (dis)similar they are. This is the distance matrix.
2. Starting with each subject as an individual cluster.
3. Merge the two nearest clusters, recording the distance between them.
4. Repeat step 3 until all subjects have been merged into a single cluster.

This converts a set of data into a series of cluster mergers along with the distances between the merged clusters (see **Figure 1** for an illustration using two-dimensional, real-valued data).

The relationship among the distances and clusters can be depicted in a dendrogram. **Figure 1** illustrates the process using artificial two-dimensional data (depicted on the left). The resulting dendrogram is depicted on the right. Each horizontal line merges two subclusters, while the height at which the horizontal line is drawn reflects the distance between the two clusters being merged. In this simple dataset, it is easy to see that subjects 1 through 5 and subjects 6 through 10 form two distinct clusters. The subjects within the clusters tend to be joined at small distances (merged at lower points in the figure), while the two distinctive clusters are further from each other (indicated by the high merge in the graph). One can also see that subjects 1 and 4 are nearest each other in the scatterplot and are merged at the lowest point in the dendrogram (at a height of approximately 0.2; enclosed in the smaller dotted box to the right). The clusters represented by subjects {6, 7, 10} and {8, 9} are further from each other, and are thus merged higher on the dendrogram (at approximately 0.9; see the larger dotted box on the left).

Clustering methods

The clustering algorithm requires a definition of "distance" between not only individual subjects, but also clusters of subjects. In the case of individual subjects, this distance is determined by step 1 above. For numerical data, some form of scaled Euclidean distance is often used (categorical data will be discussed further

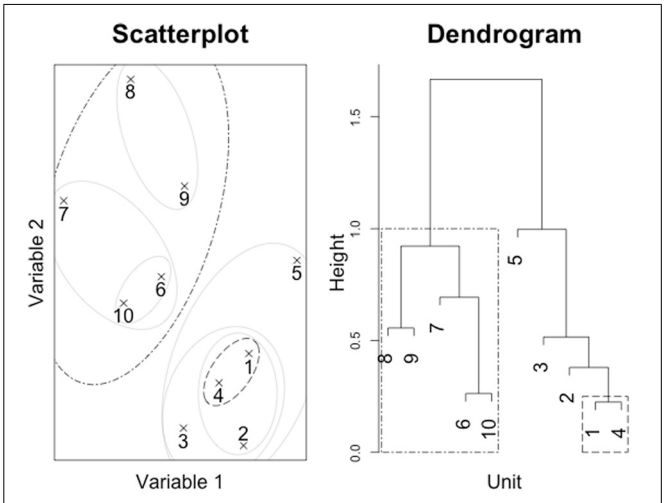


FIGURE 1 | Clustering of artificial data and dendrogram. Light gray ellipses indicate clusters, while the dark, dotted ellipses indicate the corresponding clusters in the dendrogram.

[2]Both DRC and CDP++ are in agreement on pronunciation for most non-words. Consequently, it would be difficult to draw many conclusions from a broad set of items. These 412 items amplify the differences between the models in order to better adjudicate between them.

[3]The approach described here is more accurately called agglomerative hierarchical clustering, as it starts with many clusters and ends with one. Divisive and non-hierarchical techniques are not considered.

on). However, there are many options for defining the distance between groups of subjects. The most commonly used approaches are Medoid/Centroid, Single Linkage, Complete Linkage, and Ward's method. These four methods are briefly described here, but a full treatment of clustering methods is beyond the scope of this article. The interested reader can find these techniques described in detail in cluster analysis texts such as Everitt et al. (2011). **Figure 2** depicts the results of applying them to the subject responses in Pritchard et al.'s (2012) study.

Medoid/centroid

For each cluster, the medoid or centroid is a typical element of the group. A centroid is a theoretical element that has the mean cluster value for each variable that contributes to the dissimilarity calculation. This element is likely not an actually observed element. The medoid is the individual element that is, on average, closest to all of the other elements in the cluster (in a sense the existing element that best "represents" the whole cluster).

Single linkage

The distance between two clusters A and B is defined as the smallest distance between any element in cluster A and any element in cluster B. This is sometimes referred to as the "friends-of-friends" approach, since it can result in long chains of single elements being merged into the larger cluster (**Figure 2B**).

Complete linkage

This is the complement of the single linkage approach. The distance between two clusters A and B is defined as the *largest* distance between any element in cluster A and any element in cluster B. This approach ensures that the distance between every pair of elements in the two clusters are contained within the distance between the two clusters (see **Figure 2C**).

Ward's method

Unlike the other methods described above, Ward's method does not rely on a distance metric analogous to the one used to determine the matrix in step 1 above. Instead, Ward's method minimizes the mean squared distances within the groups. At each merger, Ward's method identifies the two clusters whose merger would have the smallest influence on the mean squared within-cluster distances. Ward's method is biased toward producing spherical clusters (in essence clusters of roughly equal size; see **Figure 2D**).

The principal goal of clustering techniques is to uncover structure that may be hidden in complex data. Since this is inherently exploratory, the method that produces the most distinctive clusters in a particular data set is typically the one selected. Once clusters have been identified, a closer look at the variables that distinguish clusters from each other is necessary to determine the nature of the structure.

CLUSTERING READING ALOUD DATA

When the data being used for clustering is numerical, there are any number of approaches to defining the distance between elements. Euclidean distances between elements (using normalized variables to avoid scale effects) are common. However, in the

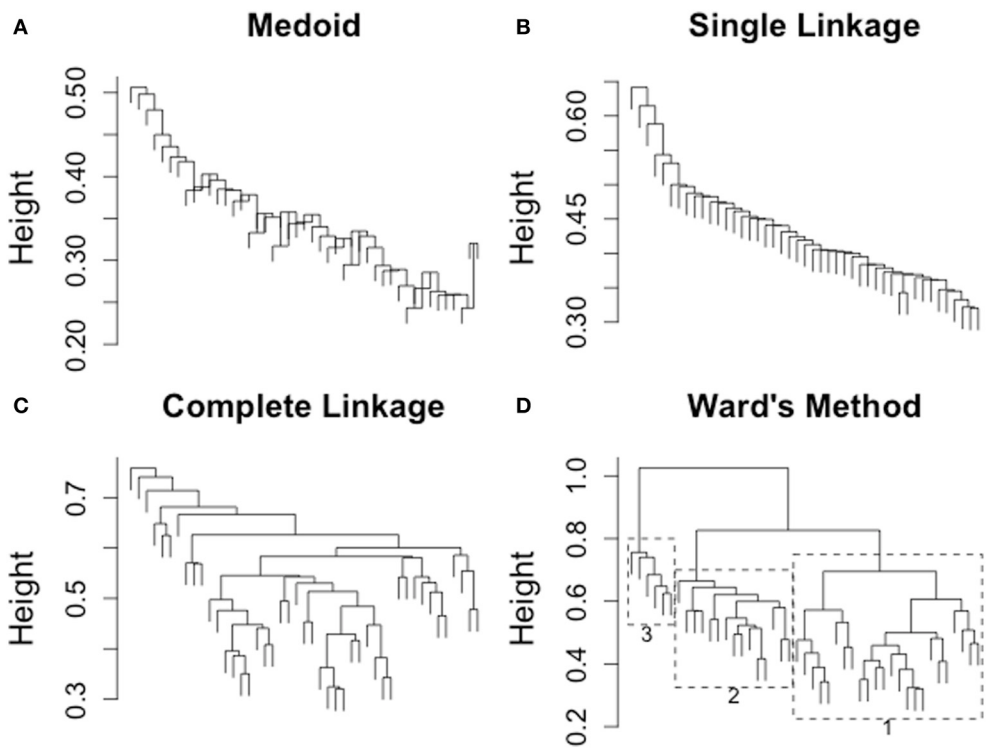


FIGURE 2 | Results of applying different clustering methods to subject data from Pritchard et al. (2012). These dendrograms do not include model responses. Methods depicted are **(A)** medoid, **(B)** single linkage, **(C)** complete linkage and **(D)** Ward's method.

Pritchard et al. (2012) study, the data are reading aloud responses to 412 non-word items. Such datasets are categorical in nature. In the case of categorical data, a pair of subjects either match or do not match on each variable. Here we opt to define the distance between two subjects as the percentage of items on which the two subjects' responses disagreed. According to this metric, a distance of 0.3 between two subjects would indicate that the subjects disagreed on 30% of the items in the Pritchard et al. dataset.

Hierarchical clustering offers us a way to simultaneously compare Pritchard et al.'s (2012) subjects across all 412 items to uncover groups of subjects that tend to be similar in their response profiles. If such latent structure can be uncovered, a closer look at the responses can help us to understand how subjects differ from each other. Further, by treating responses from computational models as theoretical subjects, we can compare the DRC and CDP++ models to the human subjects and see whether some subjects tend to cluster with one model or the other.

HUMAN READERS

Figure 2 depicts the results of clustering the Pritchard et al. (2012) data (subjects only) using each of the four clustering methods previously described. The first three methods (**Figures 2A–C**) provide little in the way of clusters for further evaluation. Ward's method (in **Figure 2D**) offers some evidence that there may be structure hidden among the subjects. Three distinct groups emerge. In **Figure 2D**, the clusters are delineated by light gray boxes and labeled in order of the size of the cluster (so that cluster 1 is the largest, and cluster 3 the smallest). Cluster 3 consists of a small subgroup of anomalous readers who are not particularly similar to each other or anyone else, while clusters 1 and 2 seem to offer more internal consistency.

Distinguishing the clusters

The power of clustering is in its ability to uncover structure that isn't based on a priori groupings (such as readers who produce regular pronunciations for non-words vs. those who produce irregular pronunciations). However, such structure is only useful if the two groups can be differentiated on the basis of their responses in some way. To determine whether and how these two groups differed, we examined the individual items by cluster and identified one possibility: the two primary groups do differ in their affinity for regularizations, but only for select ambiguous graphemes. Specifically, it seems to be a small set of ambiguous *consonant* graphemes that drive most of the difference between subject clusters.

Table 1 summarizes the types of items that underlie at least part of the difference between the two largest clusters of subjects. Consonant graphemes containing "C" are particularly discriminating, with non-words beginning with CE, CI, or CH, or ending with CE, CH, or CHE all producing different response patterns in the two clusters. "C" is not alone in discriminating clusters, however, as non-words beginning with "PH," beginning or ending with "GN," and non-words using "Y" as the only vowel cluster also discriminated. Some general observations follow.

Non-words beginning with CE or CI. For these items, subjects in cluster 1 strongly preferred to pronounce "C" with the regular /s/ over other pronunciations (85% of trials), while subjects in cluster 2 split their responses between /s/ (53%) and /k/ (43%).

Non-words ending with CE. Here again, cluster 1 subjects showed a slightly stronger preference for the regular /s/ than did subjects in cluster 2 (84% vs. 75%). What is noteworthy for these non-words is what the subjects in each cluster chose as an alternative to the /s/: cluster 1 subjects chose to infuse the item with some Italian flavor and used /tʃ/ (10%) while subjects in cluster 2 again preferred the /k/ alternative (10%). Subjects also often read these items as disyllabic (e.g., reading CICE as /siːtʃiː/). This change in syllabic parsing did not discriminate the clusters.

Non-words beginning with CH. Cluster 1 subjects again preferred the regular /tʃ/ pronunciation here (81%), or alternately a softer /ʃ/ (12%). Cluster 2 subjects also chose the regular pronunciation 67% of the time, but they were much more likely to choose an alternate, either /k/ (15%) or, less commonly, /s/ (8%). Though there is a difference in the tendency to regularize, the distinction between clusters here seems to be in the alternative pronunciations chosen.

Non-words ending with CH. Here both clusters tended to strongly prefer the regular /tʃ/ pronunciation (81% for cluster 1 and 72% for cluster 2). Again, the difference between clusters is highlighted by the alternative pronunciations with cluster 2 subjects more likely than cluster 1 subjects to choose /k/ (14% vs. 6%).

Non-words ending in CHE. For these items, cluster 1 subjects preferred the regular /ʃ/ 66% of the time, opting for /tʃ/ 27% of the time. Cluster 2 subjects showed the opposite pattern, opting for the regular pronunciation only 36% of the time, and preferring the irregular /tʃ/ 59% of the time.

For some items, cluster 2 subjects seemed to have a preference for simplifying complex or unusual graphemes. Three examples that discriminated the clusters follow.

Non-words beginning with GN. Cluster 1 subjects strongly preferred the regular /n/ for this grapheme (79% of trials), only splitting the letters into two graphemes 18% of the time (producing either /gn/ or /gən/). Cluster 2 subjects split the graphemes much more frequently (36% of trials).

Non-words beginning with PH. Cluster 1 subjects nearly uniformly chose the regular /f/ for this grapheme (99% of trials), while cluster 2 subjects occasionally seemed to ignore the H or treat it as silent, and produced /p/ on 11% of trials.

Non-words ending with GN. Here, cluster 2 subjects frequently produced responses more consistent with reversing the final phoneme. That is, they chose to pronounce the final phoneme as /ŋ/,/ndʒ/, or /ŋg/ 32% of the time rather than as the regular /n/.

Finally, Y was the only vowel that distinguished between the clusters, though not in a simple "regular vs. irregular" way.

Non-words with Y as the vowel. Both clusters were equally likely to choose the regular /ɪ/ (approximately 56% of trials). However,

Table 1 | Pronunciations that distinguished between subject clusters 1 and 2.

Pronunciation of C in CE- Items				
Cluster	***s***	***k***	***tʃ***	
1	79.7	14.0	6.3	
2	46.1	52.8	1.1	
3	44.4	55.6	0.0	
ITEMS: CERM CEBB CELK CES CEB CESH				
Pronunciation of C in -CE Items				
Cluster	***s***	***k***	***tʃ***	***ʃ***	**Other**
1	83.9	3.4	9.9	2.5	0.3
2	75.4	10.1	5.5	7.0	2.0
3	69.9	6.0	7.2	8.4	8.4
ITEMS: LARCE HACE PHLAUCE WAICE BLAUCE SKARCE WAUCE PHLEUCE CICE					
Pronunciation of C in Cl- Items				
Cluster	***s***	***k***	***tʃ***	**Other**
1	91.4	5.7	2.9	0.0
2	61.4	34.1	2.3	2.3
3	50.0	44.4	0.0	5.6
ITEMS: CICE CILTH CID				
Pronunciation of PH in PH- Items				
Cluster	***f***	***p***		
1	99.0	1.0		
2	88.8	11.2		
3	97.2	2.8		
ITEMS: PHLAUCE PHOMP PHLOMB PHRALPH PHOL PHONK PHOLK PHLEUCE PHOIN PHLOSE PHUGE PHLOTH PHUISE PHROOK PHLERSE PHLOLT PHEASE PHOZ				
Pronunciation of CH in CH- Items					
Cluster	***tʃ***	***k***	***ʃ***	***s***	**Other**
1	80.9	4.3	12.2	1.7	0.9
2	67.1	15.1	5.5	8.2	4.1
3	64.3	25.0	3.6	0.0	7.1
ITEMS: CHONGE CHIEL CHYNCH CHUILT CHACH					
Pronunciation of CH in -CH Items				
Cluster	***tʃ***	***k***	***s***	**Other**
1	80.8	6.4	10.5	2.3
2	72.4	14.1	10.6	2.9
3	64.3	18.6	13.2	3.9
ITEMS: ELCH THWONCH SMYNCH GRACH JEICH PSAUNCH GHELCH GEECH CHYNCH FRECH GYNCH THETCH STAITCH KNOUCH PSICH CHACH BLYNCH NACH GRELCH THANCH WEICH SPLACH				
Pronunciation of CH in -CHE items				
Cluster	***ʃ***	***tʃ***	***k***	**Other**
1	66.1	27.3	6.1	0.5
2	35.8	59.1	4.3	0.8
3	41.1	43.0	8.4	7.5
ITEMS: ROUCHE BOUCHE PLAUCHE DECHE THECHE DAUCHE SNICHE SKECHE SHECHE WHAUCHE VACHE BLAUCHE WRICHE FROCHE SPLICHE DRICHE SMOCHE CRICHE				

(Continued)

Table 1 | Continued

Pronunciation of Y in -Y- Items				
Cluster	**I**	**i**	**al**	**Other**
1	57.2	22.2	19.8	0.8
2	55.3	30.4	13.0	1.2
3	27.7	63.1	6.2	3.1
ITEMS: SMYNCH SCRYM NYTH CHYNCH SLYS GYNCH SMYS SMYNC FRYMPH BLYNCH GNYTH				
Pronunciation of GN in GN- Items				
Cluster	***n***	**[gn]/[gən]**	***g***	***kn***
1	79.3	18.1	1.6	1.1
2	54.7	35.9	8.5	0.9
3	27.1	50.0	18.8	4.2
ITEMS: GNANC GNEUTH GNOOSH GNUSE GNOMB GNALPH GNOSE GNYTH				
Pronunciation of GN in -GN Items				
Cluster	***n***	**[gn]/[gən]**	**[ŋ]/[ndʒ]/[ŋg]**	**Other**
1	84.5	5.6	8.5	1.4
2	56.8	9.1	31.8	2.3
3	33.3	16.7	33.3	16.7
ITEMS: VIGN BLIGN GHIGN				

if subjects chose an alternate response, cluster 2 subjects were slightly more likely to choose /i/ (30% vs. 22%) while cluster 1 subjects were more likely to choose /aɪ/ (20 vs. 13%).

DISCUSSION

It is tempting to characterize cluster 1 and 2 subjects as "regularizers" and "non-regularizers," respectively. To some extent, this may be a fair classification, but it is tempered somewhat by observations with other graphemes. First, it is noteworthy that the differences between clusters 1 and 2 do not involve *vowel* pronunciations. This is surprising as most discussion of irregularity tends to be weighted toward vowel clusters since these are generally less consistent in their pronunciations (e.g., Andrews and Scarratt, 1998; Jared, 2002). The Pritchard et al. data are consistent with the view that vowels are important to differences in responses, in that many alternate responses differed in the vowels. What the present analysis suggests is that subjects aren't naturally grouped by their vowel pronunciations. Even in the one exception to this observation (when Y is the vowel), they are not distinguished along regular/irregular lines, but rather by their choice of irregularization. To the best of our knowledge, no one has specifically examined irregularity in consonant pronunciations.

It is also not the case that the clusters can be characterized as "consonant-regular" vs. "consonant-irregular." Many ambiguous consonant graphemes do not distinguish between the two clusters at all. **Table 2** summarizes several other consonant graphemes where both cluster 1 and cluster 2 subjects showed similar patterns of regularization. That is, cluster 1 subjects are only regularizers with respect to some graphemes and not others. For example, when considering the grapheme PS at the beginning of

Table 2 | Pronunciation of other ambiguous consonant clusters that might be thought to distinguish clusters 1 and 2, but do not.

Pronunciation of SC in SC- Items			
Cluster	***sk***	***s***	ʃ
1	91.4	6.5	2.2
2	92.4	6.4	1.2
3	90.3	9.7	0.0
ITEMS: SCRUKE SCRYM SCAQUE SCROLK SCRIPE SCRALL SCROSE SCUTE SCINE SCROME SCILTH SCRALK			

Pronunciation of PH in -PH Items				
Cluster	***f***	***pf***	***p***	***v***
1	74.5	21.3	4.3	0.0
2	74.6	16.9	6.8	1.7
3	75.0	20.8	4.2	0.0
ITEMS: FRYMPH TWALPH GNALPH ZALPH PHRALPH				

Pronunciation of PS in PS- Items				
Cluster	***s***	**[ps]/[pəs]**	***sp***	**Other**
1	73.7	23.9	0.5	1.9
2	70.2	21.4	3.2	5.2
3	16.5	40.8	4.9	37.9
ITEMS: PSOOSH PSAUNCH PSAWP PSORB PSIRP PSEUCE PSAUGE PSICH PSIZ PSAR PSAISE PSAMB PSONGE PSOATH PSOOTH PSEEF PSEN PSELSE				

Pronunciation of NG in -NGE Items					
Cluster	**ndʒ**	ŋ	ŋdʒ	ŋg	**Other**
1	89.4	5.6	0.7	2.0	2.2
2	91.4	2.5	0.0	4.0	2.1
3	76.9	6.7	1.5	6.0	9.0
ITEMS: STRONGE CHONGE DONGE THWINGE SHRUNGE ENGE NENGE RENGE SNENGE SNONGE PLENGE KUNGE RINGE FRONGE YOUNGE RHINGE ZENGE PSONGE PLANGE SWOUNGE WROUNGE DANGE THINGE					

Pronunciation of TH in TH- Items				
Cluster	**θ**	***t***	**ð**	**Other**
1	97.6	1.5	0.5	0.4
2	96.8	3.1	0.0	0.2
3	93.7	4.2	0.4	1.7
ITEMS: THAC THEEL THAQUE THWONCH THEDGE THECHE THOLVE THUBE THWUILT THEIL THITE THWAZZ THUSE THRANC THODD THALC THWALC THWINGE THET THESS THAG THELM THETCH THROUSE THELK THAK THWOLVE THWELVE THWOWN THRALC THEL THRUME THREAR THWOS THANCH THESK THERP THWEB THINGE THUPE				

Pronunciation of TH in -TH Items				
Cluster	**θ**	***t***	**ð**	**Other**
1	97.6	0.0	2.2	0.3
2	97.0	0.4	1.7	0.9
3	91.6	0.0	1.1	7.4
ITEMS: SHOWTH NYTH GNEUTH STRATH FATH COWTH CILTH LOOTH SPEWTH SMOUTH PHLOTH WREWTH SCILTH PSOATH PSOOTH GNYTH				

words, they are just as likely as cluster 2 subjects to choose similar irregular pronunciations.

We turn now to a different application of the clustering algorithm. In this second analysis, we ask whether DRC and CDP++ models are better at fitting some subjects over others. Since DRC is, unsurprisingly, highly regular in its pronunciations it comes as no surprise that we would expect it to fit better with subjects from cluster 1 than from the other clusters. CDP++, on the other hand, may be better able to model subjects that tend to choose alternative pronunciations.

COMPUTATIONAL MODELS AND HUMAN READERS

Pritchard et al. (2012) compared DRC, several versions of CDP+, and CDP++ to the human response sets. The various versions of CDP+/++ tended to have very high agreement with each other. Since including several versions of CDP+/++ would induce an artificial cluster, the most successful version of the model (CDP++) could find its results dragged down by the poorer performance of the other models that it resembles. Since CDP++ had the most success in Pritchard et al.'s (2012) analysis, we include it without its siblings in our clustering analysis. This should give CDP++ the best chance of success.

The results from this clustering analysis are depicted in **Figure 3**. Subjects are labeled according to their cluster assignment from the previous analysis (excluding the models). This analysis produces two important conclusions. First, it confirms Pritchard et al.'s (2012) finding that DRC matches the responses of subjects more closely than CDP++. In the case of the clustering analysis, DRC is merged into the largest, and most homogeneous cluster of subjects (cluster 1). This suggests that DRC does an effective job of capturing the responses of a large number of subjects, allowing for some variability within and between subjects. Unsurprisingly, these are the subjects that tended toward regular pronunciations of those graphemes that distinguished the cluster 1 from cluster 2 above. It is also worth noting that DRC is merged at the lowest point in the graph. This means that no two

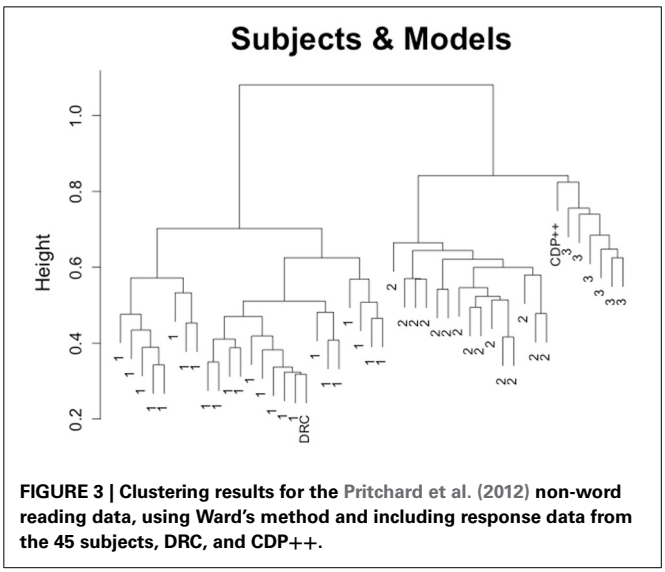


FIGURE 3 | Clustering results for the Pritchard et al. (2012) non-word reading data, using Ward's method and including response data from the 45 subjects, DRC, and CDP++.

subjects are more similar to each other than DRC is to at least one subject[4]. Second, not only is CDP++ underperforming DRC, it performs quite poorly in general, failing to match response profiles with *any* subjects and being relegated to a small group of "hermit" readers who also do not match well with any other subjects (indicated by the relatively high merge distances between and among them)[5].

DISCUSSION

DRC and CDP++ are both dual-route models, and thus share many similarities. They also both perform generally well across a range of empirical benchmarks. Adjudicating between the two models now involves closer scrutiny of individual benchmarks, and it appears that each model has a relative advantage over the other. When adjudicating between CDP++ and DRC, it would seem that different analyses arrive at different conclusions. When considering mean reaction time and accuracy data, CDP+/++ enjoys a distinct advantage over DRC because of its ability to simulate consistency effects. CDP+/++ also captures more item-level variance for words (Perry et al., 2007, 2010). However, when comparing responses directly to those produced by subjects, DRC has the upper hand. It's not clear what is at the root of this dissociation. It could be that DRC needs a more flexible set of rules and more fluidity in the possible responses in order to capture more effects and more of the item-level variance. Similarly, it may be that adjustments to CDP++'s training algorithm would allow it to learn a set of associations that more closely reflects those that subjects adopt. As things stand now, neither is clearly dominant across all of the important benchmarks for the computational modeling of reading aloud behavior.

CONCLUSION

Hierarchical clustering offers researchers a way to compare subject profiles across a range of variables. In the present study, we illustrate how hierarchical clustering of the reading aloud data from Pritchard et al. (2012) can answer two questions: first, we identified two groups of subjects who differed in their pronunciation patterns. Further, *post-hoc* examination of these clusters identified a few select consonant graphemes that underlie the difference. Critically, the differences did not conform cleanly to "regular vs. irregular" divisions. Second, we were able to provide converging evidence that DRC tends to match subjects better than CDP++. Importantly, we extend those conclusions in two ways: first DRC cannot improve much as a model of a typical skilled reader, since it fits other subjects at least as well as other subjects fit one other. In other words, the heterogeneity among subjects can never be captured by a model of an average reader that does not simulate individual differences between readers. Second,CDP++ does not appear to match any of Pritchard et al.'s 45 subjects very well, challenging a critical component of the model. The inclusion of learning algorithms to broaden a model's scope from simulating skilled reading to simulating reading acquisition may well be an important step forward (Perry et al., 2007), but CDP++ does not appear to be learning what human readers learn. Though no explicit learning algorithms are included in DRC, it appears that the rule system embedded in the GPC sublexical system better captures what skilled readers have learned about the relationship between letters and sounds.

[4]In a separate analysis we included a perfectly "modal" model. That is a hypothetical subject that always gave the modal response to each item. This model was clustered with cluster 1 and DRC, and tended to fit a few subjects better than DRC fit any subjects. That is, some subjects do appear to be more "typical" than DRC is, but DRC still performed quite well.

[5]Note that when we say that CDP++ performs poorly relative to DRC, we mean on these items. These items were specifically selected to amplify the differences between the two models so that we could more closely examine the assumptions that underlie the two models.

ACKNOWLEDGMENTS

Supported by the Australian Research Council Centre of Excellence in Cognition and its Disorders (http://www.ccd.edu.au) (CE110001021).

REFERENCES

Andrews, S., and Scarratt, D. R. (1998). Rule and analogy mechanisms in reading nonwords: hough dou peapel rede gnew wirds? *J. Exp. Psychol. Hum. Percept. Perform.* 24, 1052–1086. doi: 10.1037/0096-1523.24.4.1052

Coltheart, M., Rastle, K., Perry, C., Langdon, R., and Ziegler, J. C. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychol. Rev.* 108, 204–256. doi: 10.1037/0033-295X.108.1.204

Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster Analysis, 5th Edn.* London: John Wiley & Sons, Ltd. doi: 10.1002/9780470977811

Jared, D. (2002). Spelling-sound consistency and regularity effects in word naming. *J. Mem. Lang.* 46, 723–750. doi: 10.1006/jmla.2001.2827

Perry, C., Ziegler, J. C., and Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: the CDP+ model of reading aloud. *Psychol. Rev.* 114, 273–315. doi: 10.1037/0033-295X.114.2.273

Perry, C., Ziegler, J. C., and Zorzi, M. (2010). Beyond single syllables: large-scale modeling of reading aloud with the connectionist dual process (CDP++) model. *Cogn. Psychol.* 61, 106–151. doi: 10.1016/j.cogpsych.2010.04.001

Pritchard, S. C., Coltheart, M., Palethorpe, S., and Castles, A. (2012). Nonword reading: comparing dual-route cascaded and connectionist dual-process models with human data. *J. Exp. Psychol. Hum. Percept. Perform.* 38, 1268–1288. doi: 10.1037/a0026703

Rastle, K., and Coltheart, M. (1999). Serial and strategic effects in reading aloud. *J. Exp. Psychol. Hum. Percept. Perform.* 25, 482–503. doi: 10.1037/0096-1523.25.2.482

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.

Received: 28 January 2014; accepted: 11 March 2014; published online: 31 March 2014.
*Citation: Robidoux S and Pritchard SC (2014) Hierarchical clustering analysis of reading aloud data: a new technique for evaluating the performance of computational models. Front. Psychol. **5**:267. doi: 10.3389/fpsyg.2014.00267*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 07 November 2014
doi: 10.3389/fpsyg.2014.01229

Relative clause reading in hearing impairment: different profiles of syntactic impairment

Ronit Szterman and Naama Friedmann*

Language and Brain Lab, Sagol School of Neuroscience and School of Education, Tel Aviv University, Tel Aviv, Israel

Edited by:
Davide Crepaldi, University of Milano-Bicocca, Italy

Reviewed by:
Jon Andoni Dunabeitia, Basque Center on Cognition, Brain and Language, Spain
Carlo Geraci, Centre National de la Recherche Scientifique, Institut Jean-Nicod, France
Esther Ruigendijk, Carl von Ossietzky University Oldenburg, Germany

***Correspondence:**
Naama Friedmann, Language and Brain Lab, Sagol School of Neuroscience and School of Education, Tel Aviv University, Tel Aviv 69978, Israel
e-mail: naamafr@post.tau.ac.il

Children with hearing impairment show difficulties in sentences derived by Wh-movement, such as relative clauses and Wh-questions. This study examines the nature of this deficit in 48 hearing impaired children aged 9–12 years and 38 hearing controls. The task involved reading aloud and paraphrasing of object relatives that include a noun-verb heterophonic homograph. The correct pronunciation of the homograph in these sentences depended upon the correct construction of the syntactic structure of the sentence. An analysis of the reading and paraphrasing of each participant exposed two different patterns of syntactic impairment. Some hearing-impaired children paraphrased the object relatives incorrectly but could still read the homograph, indicating impaired assignment of thematic roles alongside good syntactic structure building; other hearing-impaired children could neither read the homograph nor paraphrase the sentence, indicating a structural deficit in the syntactic tree. Further testing of these children confirmed the different impairments: some are impaired only in Wh-movement, whereas others have CP impairment. The syntactic impairment correlated with whether or not a hearing device was fitted by the age of 1 year, but not with the type of hearing device or the depth of hearing loss: children who had a hearing device fitted during the first year of life had better syntactic abilities than children whose hearing devices were fitted later.

Keywords: hearing impairment, Hebrew, movement, reading, relative clauses, syntax, syntactic impairment, syntactic tree

INTRODUCTION

Children with hearing impairment encounter difficulties in understanding non-canonical sentences that are derived by movement of phrases (Berent, 1988, 1996a,b; De Villiers et al., 1994; Friedmann and Szterman, 2006, 2011; Friedmann et al., 2010). This deficit probably stems from limited language input during the critical period for the acquisition of the syntax of a first language (Yoshinaga-Itano and Apuzzo, 1998a,b; Mayberry et al., 2002, 2011; Yoshinaga-Itano, 2003; Friedmann and Szterman, 2006).

The aim of the current study was to learn about the nature of the syntactic deficit of children with hearing impairment. To do so, we used a novel task that allowed us to evaluate various sources for the syntactic difficulty. The task also allowed us to examine whether they experience comprehension difficulties also when the sentences are written and presented for an unlimited time, and provided a window to the reading comprehension difficulties often reported for hearing impaired children.

Studies that assessed the syntactic abilities of English-, Hebrew-, Palestinian Arabic-, and Italian-speaking hearing impaired children found difficulties in the comprehension and production of object relative clauses (English: Quigley et al., 1974a; Berent, 1988; De Villiers, 1988, Hebrew: Szterman and Friedmann, 2003, 2007; Friedmann and Szterman, 2006, 2011; Friedmann et al., 2010, Arabic: Haddad-Hanna and Friedmann, 2009; Friedmann et al., 2010; Friedmann and Haddad-Hanna, 2014; and Italian: Volpato and Adani, 2009), in the comprehension and production of object questions (English: Quigley et al., 1974b; Berent, 1996b, Hebrew: Nave et al., 2009; Friedmann and Szterman, 2011; Szterman and Friedmann, 2014, Standard Arabic and Palestinian Arabic: Friedmann et al., 2010; Haddad-Hanna and Friedmann, 2014), and in the comprehension of topicalization structures (Hebrew: Friedmann and Szterman, 2006, Arabic: Haddad-Hanna and Friedmann, 2009; Friedmann and Haddad-Hanna, 2014).

A look at these three impaired structures: object relative clauses, object questions, and topicalization structures suggests a common syntactic characteristic: they are all derived by movement of a phrase that results in a non-canonical order of the arguments in the sentence, as shown in examples (1)–(3) (movement is depicted in these examples by arrows).

(1) Object relative: This is the girl$_1$ that the grandma drew t$_1$.

(2) Object question: Which girl$_1$ did the grandma draw t$_1$?

(3) Topicalization: This girl$_1$, the grandma drew t$_1$.

In every sentence, the verb identifies the roles of the participants in the event, and assigns thematic roles to its arguments. In sentences 1–3, the verb *drew* assigns a thematic role of an Agent—the

person who draws, and a Theme—the object that was drawn. In Hebrew, as in English, verbs usually assign the theme role to the noun phrase (NP) that follows them. However, in sentences like 1–3, the Theme precedes, rather than follows, the verb. Linguistic theory suggests that such sentences are derived by syntactic movement to a position that is hierarchically higher (which typically appears earlier in the sentence) (Chomsky, 1981, 1986, 1995; Rizzi, 1990; this operation is termed "internal merge" in more recent frameworks, see Chomsky, 2000, 2001). The moved theme leaves a trace [marked in sentences (1)–(4) by t_1] (or a copy according to more recent linguistic frameworks) in its original position. The verb assigns the thematic role to the object position, and the moved object is linked to its trace with a "chain" of movement. Thus, to understand a sentence with syntactic movement, two operations are required: constructing a syntactic tree that includes a trace at the position from which the element has moved, and creating the chain between the trace and the moved element, to allow for the comprehension of the roles of the participants in the sentence.

(4) The girl$_1$ that the grandma drew t_1 is very kind.

For example, sentence (4) includes an object relative clause. In object relatives, the object of the relative clause (in this case, *the girl)* moves to a position earlier in the sentence[1]. When the object *the girl* moves, it leaves behind a trace in the embedded object position. Thus, to correctly understand such a sentence, the appropriate syntactic structure of the sentence should be constructed. This structure should include the moved element in the correct syntactic position, the relativizer (embedding marker), and an empty element, a trace of movement, should be placed in the correct position. This is not enough, though. In order to understand the role of the moved element, the chain should be established, namely, the link between the original position and the moved argument (illustrated by the arrow in 4).

A step-by-step description of the parsing the hearer needs to perform in order to understand who did what to whom in an object relative like (4) would be the following: upon hearing the NP *the girl*, the hearer is waiting for a thematic role for this NP. Once the word *that* is heard, the NP needs to be stored in a syntactic STM store until it can receive its role, and the search for a gap, the position from which this NP has moved, begins. When the subject NP *the grandma* arrives, it is also put into the syntactic store, until the verb finally arrives. When the verb arrives, the hearer accesses its entry in the syntactic lexicon together with the thematic roles it assigns. Then, the subject receives the thematic role of the Agent, the gap (trace) is postulated, and the moved element is re-accessed at this point. In processing terms, this is where the chain is constructed, between the moved element and the position in which it originated. One may think of this stage in processing terms as the re-activation of the correct antecedent at the gap (Nicol and Swinney, 1989). Impaired comprehension of a sentence with movement can result from a deficit in either of these operations.

In the current study we tried to determine which of these operations is responsible for the difficulty hearing impaired children have with object relatives. We made a distinction between the steps that require constructing the syntactic structure of the object relative clause, including the assumption of a trace, and operations related to the identification of the thematic role of the moved element (the reactivation of the appropriate NP in the trace position and the assignment of the Theme role to it).

We used a task that allowed us to separately evaluate structure building and thematic role assignment to a moved NP. This task was already used to identify the source of the deficit in the comprehension of movement-derived sentences in children with syntactic SLI and in individuals with agrammatic aphasia (Friedmann et al., 2006; Friedmann and Novogrodsky, 2007). This task used the fact that the correct pronunciation of noun-verb heterophonic homographs (i.e., words that are written the same but sound differently, like *dove*) in oral reading requires the analysis of the syntactic position of the homograph. For example, in sentence (5), the word *dove* appears as the object, and is therefore read as a noun (/dʌv/), whereas in sentence (6) it appears as the main verb, and therefore read as a verb (/doʊv/).

(5) We saw a dove flying in the sky.
(6) The dolphin dove into the river.

The dependency between correct reading aloud and the construction of the syntactic structure of the sentence served us to evaluate the way children with hearing impairment process relative clauses. We asked the participants to read aloud object relatives in which a noun-verb heterophonic homographs appeared immediately after the trace position. Thus, to read the homograph correctly in these sentences, the reader would have to be able to construct a trace after the verb, at the object position. For example, to read correctly the homograph *presents* in sentence (7), the reader has to know that the object of *received* is the trace of *the chart*, and therefore *presents* cannot be the object of *received*, and is rather the main verb. However, if the trace is not identified, the embedded verb *received* might be missing an object, so the homograph might be read as a noun, the object of *received*.

(7) The column chart$_1$ [that the scientist received t_1] presents the reading of the two groups.

Hebrew orthography allows for many degrees of freedom in the conversion of graphemes to phonemes: not all the vowels are represented in writing, some consonant letters are phonologically ambiguous, and the stress position is not marked (Friedmann and Lukov, 2008). This creates many heterophonic homographs, and for many of them one reading is a noun and the other is a verb.

[1]Some analyses assume that relative clauses are constructed by a movement of the head NP from inside the embedded relative clause CP to the relative head position above CP (raising analysis). Other analyses (matching analyses) suggest that it is an empty operator that moves from within the embedded sentence. It moves to the specifier position of CP, where it is co-indexed with the head of the relative clause (see Vergnaud, 1974; Chomsky, 1986, 1995; Kayne, 1994; and see Sauerland, 2000, for a discussion of the two analyses). For the purpose of the current study, the differences between these two approaches are irrelevant.

Many of these homographs can be used in a study of children's comprehension, because both their meanings are well-known to children.

The word MXBRT (מחברת), for example, can be read, because of the underrepresentation of vowels in Hebrew, either as a noun, /maxberet/, notebook, or as a verb, /mexaberet/, creates-feminine-3rd person-singular. Example (8) shows a sentence we used, in which the reader needs to parse the sentence and identify the syntactic role of this homograph in order to read it in a way that is appropriate for the sentence. In (8), the homograph MXBRT functions as the main verb, and is located immediately after the trace position.

(8) Ha-ganenet she-ha-yalda ohevet t_1 MXBRT sipurim.
The-kindergarten-teacher$_1$ *that-the-girl loves* t_1 *writes/a-notebook-of stories*[2].
(9) Correct reading:
The kindergarten teacher who the girl loves writes stories.
(10) Incorrect reading:
The kindergarten teacher who the girl loves a notebook-of stories.

The rationale behind this task is that if the reader postulates a trace immediately after the verb, he should know that the trace is the complement of the embedded verb *loves*. Therefore, he would analyze the homograph as the main verb, resulting with a correct reading of the homograph, as a verb (example 9). However, if the reader cannot construct a trace at the required position, the embedded verb *loves* would appear to be lacking a complement. Because the reader knows the argument structure requirements of the verb *loves*, which requires a Theme as a complement, he will search for a theme. This might lead to an incorrect reading of the homograph as the complement of the embedded verb. In this case, the written sentence (8) will be read incorrectly as in (10) *loves a notebook of stories*, where the homograph would be read as a noun. The ungrammaticality of such a reading results from the fact that the verb *loves* can only assign one thematic role of a Theme, and if the reader takes the NP after the verb to be its object and receive a Theme role, the moved element remains role-less.

The crucial point here is that even the assumption of an empty category at the correct structural position, which is enough for the correct reading aloud of the homograph, does not guarantee the correct interpretation of the sentence. If the assignment of thematic roles to the displaced NP is impaired because of a failure to establish the chain between it and its original position, the interpretation of the sentence might still be flawed. For example, an inability to assign the thematic role to the moved NP in sentence (9) might result in understanding the sentence with reversed roles, as if the kindergarten teacher loves the girl. In processing terms (see for example Nicol and Swinney, 1989; Zurif et al., 1993), this might be a result of the activation of an incorrect NP at the gap position, or not knowing which of the NPs to re-activate. Such difficulties in assignment of thematic roles can be identified by asking the reader to paraphrase the sentence.

[2]The hyphens between two morphemes or two words in the Hebrew examples indicate that they form a single written word in Hebrew.

Thus, oral reading of the homograph placed immediately after the trace position can serve as a sensitive indicator for the construction of the syntactic position of the moved element and the postulation of an empty category in its original position. The paraphrasing of the sentence can serve as an indicator for whether or not the thematic roles were correctly assigned to the moved element (and the rest of the NPs in the sentence).

If the difficulties in the comprehension of object relatives in hearing impaired children result from the inability to construct the syntactic structure and the trace, poor performance in the reading task is expected, with a tendency to read the homographic verb as the object noun. If, however, the difficulties are a result of a deficit in thematic role assignment to moved elements, with unimpaired trace identification and with good structure-building, correct reading of the homograph is expected, accompanied with difficulties in the assignment of thematic roles in the paraphrasing task. Thus, the assessment of the performance of hearing impaired children in reading and in paraphrasing of such sentences can shed light on the source of their impairment in sentences with syntactic movement.

The task can also shed light on a further open issue in the study of hearing impaired children: it is often mentioned that hearing impaired children have considerable difficulties in reading comprehension (Trybus and Karchmer, 1977; Moog and Geers, 1985; Allen, 1986; Musselman, 2000; Traxler, 2000; Moeller et al., 2006; Luckner and Handley, 2008). It might be that their reading comprehension difficulty is actually unrelated to reading, but rather stems from their syntactic difficulties. The pattern of these children's reading and comprehension of the written relative clauses (in comparison with simple sentences) might give us a further hint as to this issue.

METHOD

PARTICIPANTS

The participants were 48 Hebrew-speaking children with hearing impairment. They were 27 boys and 21 girls, aged 9;1 and 12;6 years ($M = 10;7$, $SD = 0;10$). They had moderate to severe hearing loss and were trained in oral language. At the time of testing, they were studying in primary schools in hearing classes with inclusive schooling using oral education, and each of them received additional support from a special teacher of the deaf, 2–4 h a week. All the participants consistently wore binaural hearing aids (23 children) or used cochlear implants (25 children, one of them in combination with a hearing aid on the other ear), and they all passed a hearing screening test that they performed while wearing their hearing aids/ implants, in which they were asked to repeat 10 sentences that included sibilants and were read to them by the experimenter with her lips concealed. Forty six of the participants had hearing loss from birth (based on early detection or genetic source of the hearing loss) and two had probable progressive hearing loss.

The background information on the participants' hearing is presented on **Table 1**. Subject files included no other disabilities, and in all cases neither parent was deaf, and they all came from a family that spoke only Hebrew. An informed consent statement approved by the Ministry of Education Review Board was signed by all participants' parents.

Table 1 | Background information on the hearing impaired participants.

no.	Participant	Age	Gender	Age at diagnosis	Age at the beginning of intervention (hearing aids fitted)	Type of hearing loss	Etiology	Hearing loss dB (right and left)[a]	Device (CI = cochlear implant, HA = 2 hearing aids)	Age at first implantation
1	DOH	10;10	Male	0;6	1;0	Sensorineural	Unknown	r-90, l-70	HA	
2	DOD	11;5	Female	2;6	3;6	Sensorineural	Unknown	r-60, l-65	HA	
3	CEB	9;7	Female	0;0	0;2	Sensorineural	Genetic	r-65, l-70	HA	
4	TBN	9;8	Male	0;0	0;6	Sensorineural	Genetic	r-50, l-50	HA	
5	SIG	10;6	Female	3;0	7;0	Sensorineural	Unknown	r-85, l-75	HA	
6	SAV	11;11	Male	0;6	3;0	Sensorineural	Unknown	r-45, l-50	HA	
7	AVC	10;4	Male	0;0		Sensorineural	Genetic	r-85, l-85	HA	
8	IVL	9;8	Male	0;0	3;6	Combined	Middle ear deformation	r-50, l-50	HA	
9	ORC	9;9	Male	3;0	7;0	Sensorineural	Unknown	r-65, l-120	HA	
10	TOS	10;10	Male	1;4	2;6	Combined	Genetic	r-80, l-80	HA	
11	NEA	10;3	Male	3;0	3;0	Sensorineural	Unknown	r-65, l-65	HA	
12	KEM	11;1	Female	0;6	3;0	Sensorineural	Genetic	r-70, l-75	HA	
13	ROS	10;0	Male	0;0	0;9	Sensorineural	Genetic	r-55, l-55	HA	
14	TAM	9;8	Male	0;3	0;6	Sensorineural	Preterm	r-50, l-55	HA	
15	YEO	12;0	Male	5;0		Combined	Unknown	r-50, l-55	HA	
16	DAC	10;1	Male	3;0	3;0	Combined	Unknown	r-60, l-65	HA	
17	YAO	10;1	Female	3;0	3;0	Sensorineural	Genetic	r-60, l-65	HA	
18	YIL	10;6	Female	5;0	5;0	Sensorineural	Genetic	r-80, l-80	HA	
19	LIS	11;7	Female			Sensorineural	Unknown	r-70, l-70	HA	
20	ROP	10;9	Male	0;3	1;0	Sensorineural	Genetic	r-50, l-50	HA	
21	DAM	10;1	Female	0;10	1;0	Sensorineural	Genetic	r-65, l-65	HA	
22	OFC	9;5	Male	2;0	4;0	Combined	Unknown	r-80, l-75	HA	
23	TOH	10;0	Male	0;9	0;11	Sensorineural	Unknown	r-115, l-95	HA	
24	HIM	9;11	Female	0;7	0;8	Sensorineural	Unknown		CI	1;7
25	TAC	11;3	Female	0;6	0;10	Sensorineural	Syndrome		CI	2;2
26	YOD	10;3	Male	0;6		Sensorineural	Unknown		2 CI	1;5
27	NAH	10;6	Male	0;0	0;2	Sensorineural	Syndrome		2 CI	1;0
28	SAS	10;6	Female	0;0	1;0	Sensorineural	Unknown		CI+HI	2;2
29	RON	10;9	Female	0;2	0;3	Sensorineural	Genetic		CI	1;0
30	YIB	10;3	Male	0;9	1;2	Sensorineural	Unknown		CI	1;6
31	EDY	9;6	Female	1;6	1;9	Sensorineural	Genetic		CI	4;5
32	LIH	9;1	Female	0;2	0;3	Sensorineural	Unknown		CI	1;0
33	LER	9;11	Male	0;0		Sensorineural	Genetic		2 CI	1;3
34	RAR	11;7	Male	0;0	1;0	Sensorineural	Unknown		CI	5;0
35	LIW	10;1	Male	0;6		Sensorineural	Genetic		CI	1;0
36	LIA	10;0	Female	0;8		Sensorineural	Unknown		2 CI	1;3
37	KOZ	11;8	Male	0;3	0;3	Sensorineural	Genetic		CI	1;0
38	AIR	11;5	Female	1;0	1;0	Sensorineural	Genetic		2 CI	1;8
39	LIC	11;3	Female	0;8	0;8	Sensorineural	Genetic		2 CI	1;2
40	ZIZ	11;0	Male	0;9	0;9	Sensorineural	Unknown		CI	2;1
41	ORS	9;10	Male	0;11	1;0	Sensorineural	Genetic		CI	2;1
42	MAK	11;1	Male	0;9	0;9	Sensorineural	Unknown		CI	4;0
43	TOS	12;6	Male	0;10	1;0	Sensorineural	Unknown		CI	4;9
44	ODC	10;10	Female	0;0	0;3	Sensorineural	Genetic		CI	3;0
45	CBN	12;2	Male	0;0	0;6	Sensorineural	Genetic		CI	3;6
46	LIL	10;10	Female	0;0	0;4	Sensorineural	Genetic		2 CI	5;0
47	MAL	10;10	Female	0;0	0;4	Sensorineural	Genetic		2 CI	2;6
48	YIC	11;5	Female	0;0	0;6	Sensorineural	Genetic		CI	1;6

[a] *All the participants with unilateral cochlear implant had a hearing loss of 95–105 dB in the unaided ear (pure tone average of 500, 1000, and 2000 Hz).*

r, right ear; l, left ear; CI, unilateral cochlear implant; 2CI, bilateral cochlear implants.

To evaluate oral reading at the single word level, and exclude participants with severe dyslexia, we tested 43 of the hearing-impaired participants using the TILTAN screening test (Friedmann and Gvion, 2003), which was developed to identify subtypes of dyslexia. The screening test includes oral reading of 136 single words, 30 word pairs, and 40 non-words. The test includes words of various types that can reveal the different types of dyslexia. The results of the screening test indicated that two girls of the initial group of 50 participants, had a significant deficit in reading words, and therefore they were excluded from this study.

Control group

The participants in the control group were 38 typically-developing hearing Hebrew-speaking children in fourth grade (mean age = 9;8, *SD* = 0;5). They met the criteria of normal hearing, normal language development, and had no reports of neurological development difficulties or socio-emotional problems. They were taken from public schools serving a middle-class population, similarly to the participants with hearing loss.

MATERIALS

The test included 20 sentences in which the main verbs that were heterophonic-homographs of nouns. Half of the sentences were relative clauses, and half were simple control sentences. The relative clauses were center-embedded object relatives with the relativizer "she-", which is obligatory in Hebrew relative clauses (it is also used as the embedding marker for sentential complements). The homographic verbs appeared in the relative clauses immediately after the trace. Because we needed to add the homograph after the trace position, we used center-embedded object relatives. Hebrew-speaking children at the ages tested already understand such center embedded relatives, as shown in Friedmann and Novogrodsky (2007) and in the performance of the hearing control participants of the current study reported below.

Example (11) shows an object relative clause with the homograph עלה (ʔLH), which can be read either as the verb /ala/, meaning ascended, climbed, or as the noun /ale/, a leaf. For each sentence with a relative clause, a control sentence that included the same homograph was constructed that was a length-matched simple sentence without movement (12).

(11) ha-madrix$_1$ she-ha-yeled ra'a t$_1$ ala al ha-har.
The-guide$_1$ that-the-boy saw t$_1$ climbed the mountain
The guide that the boy saw climbed the mountain.

(12) ha-sus im ha-zanav ha-gadol ala al ha-deshe.
The-horse with the-tail the-big climbed on the grass
The horse with the big tail stepped on the grass.

The relative clauses were constructed so that the homograph in its incorrect, noun reading would be a semantically and syntactically appropriate complement of the embedded verb. For this aim we chose embedded verbs that could take the noun homographs as their object. There were no morphological cues that could identify the homograph as a verb or a noun. The fact that the homograph was not preceded by an article could also not be used as a cue for it being a verb, because in Hebrew indefinite nouns appear without any determiner. To prevent reliance on semantics and world knowledge cues in the interpretation of the sentences, the relative clauses were semantically reversible, namely, the two NPs in the sentence could semantically both serve as the agent and as the theme of the embedded verb, and both could serve as the agent of the main verb. For example, in (11) it is possible both for the guide to see the boy and for the boy to see the guide, and both the boy and the guide can climb the mountain, and therefore comprehension cannot be based solely on the semantics of the lexical items, and has to rely on syntax. The sentences included only homographs for which the verb and the noun meanings were different enough to permit reliable judgment of which meaning was selected in the speakers' paraphrases (like *dove, tear, presents*, and *objects* in English). The homographs were simple and frequent words that school-age children are acquainted both with their verb and with their noun meaning.

The homographs in the sentences were either biased toward the incorrect (noun) meaning or had two meanings with similar frequency. The dominant meaning was determined by Friedmann and Novogrodsky (2007) according to judgments of 50 Hebrew-speaking adults and 50 Hebrew-speaking children without language impairment. The 50 adults (aged 18–55) were asked to determine for each heterophonic homograph which of the meanings is more frequent—they could either circle one meaning or say that they had similar frequency. According to their judgments we classified one homograph as biased toward the noun reading, and the rest as equi-biased (according to Onifer and Swinney's 1981 criterion for primary meaning, of a meaning preferred by at least 75% of the judges)[3]. In addition, Friedmann and Novogrodsky (2007) presented a list of the homographs as single words to 50 children in 4th–7th grade and asked them to read them aloud, and noted how often each homograph was read as a noun or as a verb. The results were similar to the results of the adults, and even more strongly biased toward the noun reading. Eight of the homographs were strongly biased toward the noun reading (more than 75% of the children read them as nouns), and two homographs were biased toward the noun but less strongly (69 and 74% of the children read them as nouns).

The test sentences were divided into two blocks; one block was administered in each of two sessions, in each block each homograph appeared only once. Each block included five relative clauses and five control sentences, in random order. The second block included the control sentences for the five target sentences in the first block, and five relative clauses whose control sentences appeared in the first block.

PROCEDURE

The sentences were printed on a white page, presented in front of the participants. We asked the participants to read each sentence aloud, and then to explain it in their own words. To explain what

[3]Using a definition of ambiguity bias according to which the difference between the number of judges who preferred the noun meaning and the number of judges who preferred the verb meaning was at least 25% (13) of the judges there were six words biased toward the (incorrect) noun reading, and four which were unbiased.

"in your own words" mean, we started with a simple sentence, and gave the participants feedback on their paraphrase. For the rest of the task we did not give any feedback as to their success or failure to understand the sentences, only commented on whether or not their paraphrase was complete, and gave general encouragement.

If, in the paraphrase, the child only explained part of the sentence or if the paraphrase was not sufficiently clear to determine the thematic roles the participant assigned to the NPs in the sentence, we asked a clarification question. For example, if a participant said in the paraphrase of sentence (11) "The boy saw the guide", and ignored the main verb, we asked "and what else happened in the sentence?", and when a participant said "The guide climbed the mountain", we asked "and what about the boy?". When the participants repeated the written sentence, we asked them to try again, and explain the sentence in their own words.

No time limit was set. The sentences remained in front of the participants throughout the reading and paraphrasing task, to reduce demands on memory. The two 10-sentence blocks (in which the same homographs were incorporated in different sentences) were administered 1 or 2 weeks apart.

ANALYSES

Reading aloud was classified as correct if the homograph was read correctly, immediately or after self-correction. Paraphrases were classified as correct if they described correctly the thematic roles of the two NPs in the sentence and the arguments of the two verbs—the main verb and the embedded verb. Paraphrases in which one or more thematic roles were incorrect were counted as incorrect.

RESULTS AT THE GROUP LEVEL

The results, summarized in **Figure 1**, indicate that the hearing impaired children have a severe difficulty in object relative clauses. The group's reading and paraphrasing of the object relatives were significantly poorer than that of the control group. Importantly, the difficulty exhibited by the hearing impaired group did not result from a general difficulty in reading or in paraphrasing. They performed very well in reading and paraphrasing the simple length-matched control sentences, which did not include movement, and their performance was virtually like that of the control participants in these sentences.

READING ALOUD

As a group, the hearing impaired children showed difficulty in the reading of the homograph in the object relatives, and their performance ($M = 81.5\%$, $SD = 22.6\%$) was significantly poorer than that of the participants in the control group ($M = 93.8\%$, $SD = 7.2\%$), Wald chi $= 25.41$, $p < 0.001$ (analyzed using a mixed logit model, with by-participant and by-item random effects).

Out of the 480 homographs in object relatives that the group of hearing impaired children read, only 391 were read correctly. All the 89 sentences that were read erroneously included incorrect reading of the homographic verb as a noun [see (13) for an example for the way they read the sentence with the homograph מגדל, MGDL, which can be read either as a verb /megadeal/ "grows", or as the noun /migdal/ "tower"].

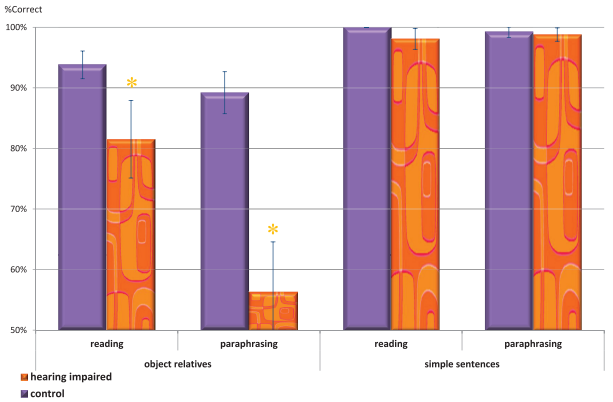


FIGURE 1 | Reading and paraphrasing of relative clauses and simple control sentences in the two groups. Error bars present 95% confidence interval. $*p < 0.001$ in the comparison between the groups.

Other important information can be gained by looking at the children's reading of other parts of the relative clause. Some of the children canceled the subordination in the sentence by changing the relativizer "she" into the word "shel", *of* in Hebrew (14), or into a coordination marker.

(13) a. Target sentence:
ha-baxur$_1$ she-aba cilem t$_1$ megadel taltalim arukim.
the-guy$_1$ that-dad photographed t$_1$ grows curls long
The guy that dad photographed grows long curls.
b. Reading the homograph verb as a noun:
ha-baxur she-aba cilem migdal taltalim arukim.
the-guy that-dad photographed tower curls long
The guy that dad photographed a tower of[4] long curls.

(14) Canceling the subordination:
ha-baxur shel aba cilem megadel taltalim arukim.
the-guy of dad photographed grows curls long
Dad's guy photographed grows long curls.

Importantly, the marked difficulty in reading the homograph cannot be ascribed to a reading impairment, but rather stems from the syntactic structure of the relative clause: When the same homographs were incorporated in simple sentences, the hearing impaired participants read them very well (98% correct), and significantly better than when they were incorporated in object relatives (81%), Wald chi = 48.32, $p < 0.001$[5].

[4]In Hebrew, compounds like "tower of curls" are two-word phrases without the word "of" (Doron and Meir, 2013).

[5]Another finding that supports the idea that impaired reading-decoding abilities do not underlie the hearing-impaired participants' failure on object relatives in this task is that individuals with developmental dyslexia (surface dyslexia, letter position dyslexia, attentional dyslexia, and visual dyslexia) without a syntactic problem who were tested on the same task still read the homograph in the relative clauses well, and similarly to their reading of the same homograph in the simple control sentences, and could interpret the relative clauses correctly (Kesselman et al., 2013).

PARAPHRASING

The paraphrasing task also indicated that as a group, the hearing impaired children have a considerable difficulty in paraphrasing the object relatives. Their performance in paraphrasing of the object relatives ($M = 56.9\%$, $SD = 29.4\%$) was significantly poorer than that of the participants in the control group ($M = 89.2\%$, $SD = 10.9\%$), Wald chi = 95.47, $p < 0.001$. This difficulty did not result from a general problem in the task of paraphrasing, as indicated by their good paraphrasing of the control sentences ($M = 98.8\%$, $SD = 3.9\%$), which was significantly better than their paraphrasing of the object relatives, Wald chi = 67.49, $p < 0.001$, and not differently from that of the controls, Wald chi = 0.08, $p = 0.77$.

Out of 480 object relatives the hearing impaired children paraphrased, only 270 sentences (56%) were paraphrased correctly. In marked contrast, when they paraphrased the simple control sentences, they did it well, and made errors only on 6 sentences out of 480 sentences (1%).

There was a main effect for sentence type, Wald chi = 67.49, $p < 0.001$, no significant main effect of group, but importantly, a significant interaction between sentence type and group, Wald chi = 6.06, $p = 0.01$.

We further analyzed the paraphrasing errors. The detailed distribution of the paraphrasing errors in the hearing impaired group is presented in **Table 2**. One of the two most common types of paraphrasing errors was incorrect thematic role assignment. The incorrect thematic role assignment errors, which accounted for 55% of the errors in paraphrasing, included three types of incorrect thematic role assignment. One was ascribing the predicate of the main clause to the subject of the relative clause, which occurred in 26% of the thematic role errors [see example (15a) for a paraphrase that one of the participants gave for sentence (15)]. Another error type in thematic role assignment involved ascribing the predicate of the relative clause to the subject of the main clause, which occurred in 33% of the thematic role errors (15b). Additional 41% of the thematic role errors in paraphrasing involved both ascribing the predicate of the main clause to the subject of the relative clause, and ascribing the predicate of the relative clause to the subject of the main clause (15c).

Another frequent type of error involved the interpretation of the homograph as a noun (16). In these paraphrases the hearing impaired children tried to make sense of the sentences somehow and to reach an interpretation in which all NPs in the sentence receive a role. Additional incorrect responses included cancelation of the subordination and "I don't understand" responses. Some responses included more than one type of error.

Table 2 | The distribution of the paraphrasing errors of the center-embedded object relatives in the hearing impaired group.

Paraphrasing error type	% of paraphrasing errors
Incorrect thematic role assignment	55.0
Treating the homograph as a noun	32.0
Cancelation of the subordination	7.5
"I don't understand" responses	5.5

(15) Target sentence
ha-baxur$_1$ she-aba cilem t$_1$ megadel taltalim arukim.
the-guy$_1$ that-dad photographed t$_1$ grows curls long
The guy that dad photographed grows long curls.
Correct paraphrasing:
Aba cilem et ha-baxur ve-ha-baxur megadel taltlim arukim.
Dad photographed ACC the-guy and the guy grows long curls
Dad photographed the guy and the guy grows long curls.

Incorrect thematic role assignment

a. **Ascribing the predicate of the main clause to the subject of the relative clause**
Aba, yesh lo…hu megadeal taltalim arukim.
Daddy, there's to-him…he grows curls long
Daddy, he has…he grows long curls.
b. **Ascribing the predicate of the relative clause to the subject of the main clause**
Ha-baxur she-cilem et aba hu megadeal taltalim.
The-guy that-photographed ACC daddy he grows curls
The guy that photographed daddy, he grows curls.
c. **Both ascribing the predicate of the main clause to the subject of the relative clause, and ascribing the predicate of the relative clause to the subject of the main clause**
Aba megadel taltalim ve-ha-baxur cilem oto
daddy grows curls and-the-guy photographed him
Daddy grows curls and-the-guy photographed him

(16) **Treating the homograph as a noun**
Aba cilem et ha-migdal taltalim arukim.
Daddy photographed ACC the-tower curls long
Daddy photographed the long curls tower.

In 121 of the sentences, the hearing impaired children paraphrased the sentences erroneously although their reading was correct. In 89 other cases the incorrect paraphrasing of the center-embedded object relatives was a result of incorrect reading of the verb-noun homograph.

RESULTS AT THE INDIVIDUAL LEVEL: CRUCIALLY DIFFERENT PROFILES

GROUP-LEVEL ANALYSIS HIDES TWO DIFFERENT PROFILES OF IMPAIRMENT

The analysis of the performance at the group level shows a significant difficulty both in reading homographs placed after the trace position, and in paraphrasing object relatives. However, the group analysis hides crucially different profiles within the hearing impaired group. When we analyzed the performance of each of the hearing impaired participants in reading and paraphrasing, we found that they do not all show the same pattern. In fact, three different patterns could be detected. One subgroup of hearing impaired children read the homographs in the object relative clauses correctly, much like the controls, but failed to explain the meaning of the object relative clauses. Another subgroup showed severe difficulties both in reading the homographs in object relatives, and in paraphrasing the object relatives. The third subgroup showed relatively normal reading and paraphrasing of relative

clauses. For this analysis, we classified each participant in the group of children with hearing impairment into the subgroups according to whether s/he failed in reading and whether s/he failed in paraphrasing compared to the hearing control group.

We defined failure in reading or paraphrasing as performance that is significantly below that of the hearing children. These comparisons of the performance of each of the experimental participants with the performance of the normative hearing control group were done using Crawford and Howell's (1998) *t*-test (see also Crawford and Garthwaite, 2002), $p < 0.05$. This test is a modification to the independent samples *t*-test that can be used to compare an individual, treated as a sample of $N = 1$, with a sample, in a way that the single participant does not contribute to the estimate of the within-group variance. This analysis according to failure in reading and/or paraphrasing created three subgroups that were confirmed also with a discriminant analysis. The two discriminant functions, using prior probabilities, predicted correctly 91.7% of the classification.

As summarized in **Table 3**, 11 children read the homographs in the relative clauses well (not significantly different from the control participants), but paraphrased them incorrectly, and significantly worse than the controls ($p < 0.05$). Their good reading of the homographs placed after the trace in object relatives indicates that these children are able to construct the syntactic structure correctly and to postulate an empty category in the trace position, and therefore they assume that the homograph is a verb, and not the object of the embedded verb. However, their failure to paraphrase the object relatives indicates that although they assumed an empty category in the right place, they were unable to establish the chain between the trace and the moved NP and hence did not assign the moved NP the correct thematic role. The good structure building of these participants goes well with previous descriptions of the Wh-movement deficit in hearing impaired children resulting from a problem in the chain of movement rather than from a deficit in the syntactic structure. For example, in a previous study (Friedmann and Szterman, 2011), we showed that the hearing impaired children in their study produced embedded sentences very well, indicating that they were able to construct the syntactic tree up to its highest node (we will explain in detail about the syntactic tree below).

Other 15 children showed poor reading of the homographs and poor comprehension of the relative clauses with the homographs (significantly poorer than the control group). Their reading indicates that it was not only their ability to assign thematic roles to the moved NP that was impaired. They did not even know that there was a movement in the sentence, and failed in the construction of the syntactic structure and the trace. How can this deficit be characterized? We suggest that it could be a structural problem in constructing the syntactic tree up to its highest nodes.

When speakers produce or comprehend sentences, they represent them in syntactic trees (Pollock, 1989; Chomsky, 1995, see **Figure 2**). The phrasal architecture of the syntactic tree consists of three main structural layers, which are, from bottom to top, the lexical layer, the inflectional layer, and the complementizer layer (Chomsky, 1986, 1995; Rizzi, 1997). The lexical layer VP (verb phrase) contains the subject, the verb, and the object; the inflectional layer IP (inflectional phrase) is responsible for verb

Table 3 | Individual profiles: number of homographs in object relatives read correctly and of object relatives paraphrased correctly out of 10 object relatives.

Participant		Syntactic impairment	Homograph reading	Paraphrase
GOOD READING, POOR PARAPHRASING				
1	DOD	Movement impairment	10	3
2	SIG	Movement impairment	9	5
3	TOS	Movement impairment	8	3
4	YEO	Movement impairment	10	3
5	YAO	Movement impairment	10	6
6	OFC	Movement impairment	8	2
7	TOH	Movement impairment	9	6
8	YOD	Movement impairment	10	6
9	ODC	Movement impairment	8	6
10	CBN	Movement impairment	9	3
11	YIC	Movement impairment	10	6
POOR READING, POOR PARAPHRASING				
1	DOH	CP impairment	7	4
2	AVC	CP impairment	4	1
3	IVL	CP impairment	7	3
4	ORC	CP impairment	4	1
5	DAC	CP impairment	7	1
6	LIS	CP impairment	7	6
7	HIM	CP impairment	8	4
8	NAH	CP impairment	5	5
9	YIB	CP impairment	5	3
10	LIH	CP impairment	7	2
11	LER	CP impairment	7	4
12	RAR	CP impairment	4	3
13	ZIZ	CP impairment	2	1
14	ORS	CP impairment	3	1
15	MAK	CP impairment	3	1
GOOD READING, GOOD PARAPHRASING				
1	CEB		10	8
2	TBN		10	8
3	SAV		10	8
4	NEA		9	8
5	KEM		10	9
6	ROS		9	8
7	TAM		10	9
8	YIL		10	9
9	ROP		10	10
10	DAM		9	9
11	TAC		10	7
12	SAS		10	7
13	RON		9	7
14	EDY		10	10
15	LIW		9	7
16	LIA		8	7
17	KOZ		10	10
18	LIC		10	10
19	TOS		10	10
20	LIL		9	9
21	MAL		9	7
22	AIR		9	7
Control group: average mean (SD)			9.4 (0.7)	8.9 (1.1)

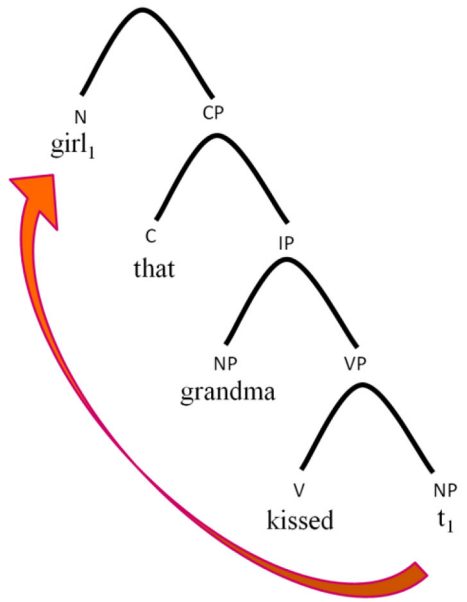


FIGURE 2 | The syntactic tree of a noun phrase with an object relative clause.

inflections; the CP (complementizer phrase) layer is responsible for embedding and for constituents that move to the beginning of the sentence such as Wh-morphemes and moving elements in relative clauses, verbs that move to second sentential position, and auxiliaries in yes/no questions in some languages. The CP-layer is the highest layer in the sentential hierarchy.

If these participants could not construct the tree up to the CP layer, which is responsible for embedding markers and Wh-movement, then they could not even know, when reading the sentence, that they need to be looking for a trace. Hence, they did not detect the trace position, which, in turn, led to their incorrect reading of the homograph. If this is a correct portrayal of their deficit, this has far-reaching predictions for the performance of this subgroup of hearing impaired children in other aspects of their linguistic performance—we would expect these children to show additional indications of CP impairment.

FURTHER ASSESSMENT OF CP IN THE SUBGROUP WITH POOR HOMOGRAPH READING

To determine whether the hearing impaired children who read the homographs incorrectly indeed have a structural deficit at the CP level, we looked at these children's pattern of errors in the current task, and also examined their performance in other sentence types and tasks that can serve as markers for the status of their CP. Their performance in both the current study and other tasks was revealing and supported the conjecture about a CP impairment.

Their reading of the relative clauses differed from that of the other hearing impaired participants not only in that they read the homograph as a noun. It also differed in another important aspect. When they read the sentences and reached the position of the embedding marker, they sometimes canceled the embedding [see example (14) above]. This problem with embedding was evinced both in their reading and in their paraphrasing of embedding. Thus, the reading pattern on the relative clause reading task, beyond the incorrect reading of the homograph, indicates a difficulty with embedding in this subgroup. This also affected the errors the children in this group made in paraphrasing the object relatives: whereas the children who were impaired in Wh-movement made mainly errors of thematic role assignment (79% of their errors) and the rest were interpreting the homograph as a noun (21%), the children with suspected CP impairment made many paraphrasing errors that stem from failed structure building: they had 54% errors of interpreting the homograph as a noun, 32% errors of thematic role assignment, and 15% paraphrases that disregarded the embedding. Importantly, each of the children in the CP group had errors of interpreting the homograph as a noun, and eight of the 15 children in the CP group showed cancelation of the embedding in their paraphrases. Thus, the children in the CP group made significantly more errors in paraphrasing that involved interpreting the homograph as a noun [$t_{(23)} = 4.49$, $p = 0.0001$], and canceling the embedding [$t_{(23)} = 1.73$, $p = 0.048$] than the children in the Wh-movement impairment group.

Findings from other tasks support the hypothesis regarding these children's inability to project the CP even in sentences that do not involve Wh-movement. We selected structures and tasks that are expected to be affected by a CP impairment, but not by a problem with the assignment of a thematic role to a moved NP that had undergone Wh-movement across another NP. We used tasks and structures that, as shown in **Table 4**, typically developing Hebrew-speaking children already perform very well in the age range of the hearing impaired participants. We had six tasks examining four different structures that corresponded to these criteria. One such structure are sentences with a subject relative clause, which include embedding but in which Wh-movement does not cross another NP. A deficit in the assignment of a thematic role to the moved NP across another NP would not affect the production of subject relatives, but a deficit in the CP layer is expected to affect subject relatives because they involve embedding and the CP layer. With a similar rationale in mind, we also looked at subject questions, which also involve the CP layer but no crossing movement. Subject questions, like subject relatives, are expected to be difficult if a child has a CP impairment but not if the impairment is constrained to Wh-movement across another NP. Two other structures with which we tested the CP layer were embedding without Wh-movement, with sentential complements of verbs (which involve the C node but no Wh movement), and sentences with verb movement to C, again, a structure that involves the CP layer but no movement of an NP across another NP.

Because the comprehension of subject relatives and subject questions can rely on the canonical words order and possibly an agent-first strategy, we tested these structures using production and repetition tasks. Subject relative production was assessed using two tasks: a preference task in which the child heard descriptions of two children and was requested to choose the child he would rather be, using a subject relative clause (BAMBI ADIF, see Friedmann and Szterman, 2006; Novogrodsky and Friedmann, 2006, for details on this task), and a picture description task

Table 4 | Performance of the children with impaired homograph reading (and hence, suspected impairment in CP) in additional CP tasks, compared with hearing controls and the children with good homograph reading.

S. no		Subject relative production: preference task	Subject relative production: picture task	Subject questions: repetition	Sentential complements of verbs: repetition	Verb movement to C: repetition	Verb movement to C: comprehension
HEARING-IMPAIRED PARTICIPANTS WITH CP DEFICIT							
1	DOH	100	70	60	Impaired[a]	80	100
2	AVC	90	80	100	100	100	100
3	IVL	100	100	100	100	60	100
4	ORC	100	100	100	100	20	100
5	DAC	100	100	100	100	40	87
6	LIS	100	40	100	100	40	100
7	HIM	100	100	80	100	60	100
8	NAH	80	70	100	100	60	–
9	YIB	100	60	100	100	0	87
10	LIH	90	80	40	100	80	100
11	LER	100	100	100	40	60	75
12	RAR	10	100	60	80	40	37
13	ZIZ	0	70	40	100	0	100
14	ORS	–	–	80	60	0	37
15	MAK	40	100	100	100	–	–
RESULTS OF THE HEARING CONTROL GROUPS							
% correct: average (SD)		100	98 (3.9)	100	100	87 (18)	97 (5.4)
Age: mean (SD)		8;0 (0;6)	9;0 (1;1)	5;10 (0;4)	5;10 (0;4)	7;9 (0;6)	10;3 (1;0)
RESULTS OF THE OTHER HEARING IMPAIRED SUBGROUPS							
% correct: average (SD), and number of participants below control							
The 22 hearing-impaired with normal performance on the homograph task		99 (2.2) 1	98 (3.7) 0	95 (11.2) 3	98 (7.1) 1	81 (23.5) 5	98 (8.7) 1
The 11 hearing-impaired with good reading and poor paraphrasing		92 (9.7) 4	86 (21.3) 3	83 (22.5) 4	100 0	78 (32.8) 2	92 (13.2) 1

Shaded cells indicate that the participant performed this task significantly below the hearing control group.

[a]DOH repeated only three sentential complements, but a different task points to his embedding difficulty: in a test of pronoun comprehension, he performed only 50% correct in the embedded sentences (whereas in the simple sentences he performed well).

that elicited subject relatives (BAMBI ZIBUV, see Friedmann and Szterman, 2006, for details on this task. We used a repetition task (PETEL repetition task, Friedmann, 2000; see Fattal et al., 2011; Friedmann and Szterman, 2011 for details on this task) to evaluate the repetition of subject questions, as well as the repetition of sentences with a clause embedded to a verb and sentences with verb movement to C. The comprehension of sentences with verb movement to C was assessed using a task that is somewhat similar to the task described in the current article. The children heard sentences with the verb in second position after an adverb, in which a pseudoword was placed in the verb position or in the object position. Understanding whether the pseudoword is a noun or a verb required the correct construction of the sentence structure, including the verb movement to C. The children were then asked what the pseudoword could mean in each sentence, and we tested whether they suggested a verb or a noun (Szterman and Friedmann, 2011).

As shown in **Table 4**, each of the children who failed in the homograph reading task showed clear indications of difficulties in these CP-related tasks, and performed poorly in at least one task that the typically developing hearing children already perform very well in their age range or even in much younger ages[6,7]. Most of them (11 of the 15) failed on 2 or more such tasks. Importantly,

[6]The results of the hearing children in **Table 4** are taken from previously reported data that used the same tests. The data on subject relative clause elicitation is from Fattal et al. (2013) for the preference task and from Novogrodsky and Friedmann (2006) for the picture task, the data on the repetition of subject questions and sentences with sentential embedding are taken from Fattal et al. (2011), the data on the repetition of sentences with verb movement to C are taken from Fattal et al. (2013) and the data on the comprehension of sentences with verb movement to C is taken from Szterman and Friedmann (2011).

[7]The poor performance on the comprehension and repetition task could not be ascribed to the participants' not hearing the sentences well. These tasks included simple control sentences on which the participants performed well, indicating that hearing the sentences was not the problem, but rather their syntactic structure. Recall, also, that all the participants passed a hearing screening test (see Participants section).

the performance of the children with suspected CP impairment was not only much poorer than that of hearing children, but also poorer than the performance of the other subgroup of hearing impaired children, the ones who read the homograph correctly but failed to paraphrase the object relatives. As shown in **Table 4**, these children performed well on the CP tasks, much better than the hearing impaired children who failed in the homograph reading[8].

Perusal of earlier literature that analyzed the performance of hearing impaired children in these tasks and structures sheds further light on the two profiles of impairment. First, it shows that other hearing impaired children who fail on Wh-movement can still perform well on tasks that involve CP but not Wh-movement. In previous studies Friedmann and Szterman (2011) assessed the comprehension and production of Wh questions and relative clauses in a different group of 18 Hebrew-speaking hearing impaired children. The production of subject and object questions was assessed using an elicitation task with 40 pictures of a figure doing something to another figure, in which the agent or the theme figure was concealed, and the child was instructed to ask a question about the concealed figure. Friedmann and Szterman found that 11 of the hearing impaired children failed to produce Wh questions in this task. Importantly, two profiles were detected in their production of Wh questions: Seven of the participants failed to produce object questions but still produced subject questions normally, whereas four participants showed difficulties in both object and subject questions. The difficulty in subject questions of these four children was also manifested in their poor comprehension and poor repetition of subject questions.

Similar results were found in a study that assessed the comprehension and production of relative clauses in Hebrew-speaking hearing impaired children (Friedmann and Szterman, 2006). Whereas 11 of the 14 participants performed better in subject relatives than in object relatives, 3 of the participants showed difficulties in both subject relatives and object relatives.

Finally, in a study of the comprehension of verb movement to C (Szterman and Friedmann, 2011), 9 of 12 hearing impaired participants performed well and similarly to the hearing control group (with no more than a single error) in the comprehension of sentences with verb movement. Three participants showed a different pattern, and failed to understand these sentences.

Thus, when previous data are analyzed at the individual participant level, they already suggest that two different patterns of syntactic impairments can be detected in hearing impaired individuals. We suggest that the 15 hearing impaired participants who failed to read the homograph in the object relatives in the current study had difficulty in the construction of the syntactic structure, presumably in the construction of the syntactic tree up to its highest node, CP. This difficulty was expressed in the way they read and paraphrased the object relatives, where they showed clear indications of difficulties in embedding, as well as in other tasks, in which they failed in the comprehension, repetition, and production of structures that did not involve Wh-movement (across another NP) but did involve CP: embedded sentences, verb movement to C, subject questions, and subject relatives. These structures are already mastered at this age by hearing children, and, importantly, by the hearing impaired children in the current study who read the homograph correctly also showed good performance in the CP tasks.

PREDICTORS OF COMPREHENSION OF MOVEMENT-DERIVED SENTENCES

Given the two general patterns that we found: impaired and unimpaired syntax, and the further division into two profiles of impairment, we tried to see whether any of the background measures—depth of hearing loss, type of hearing aid, and age of fitting of a hearing aid—could be responsible for these groupings. We could not find any factor that determined, within the group of syntactically impaired children, who will show the movement deficit and who will show the CP deficit. However, the difference between the syntactically-impaired group and the group with normal syntax did correlate with one factor: whether or not hearing devices (be they cochlear implants or hearing aids) were fitted by the age of 1 year.

The age of fitting of a hearing aid was the only background factor that correlated with syntactic performance: Phi Coefficient of Association, calculated for the age of intervention (before or after 1 year) and performance (intact or impaired, determined using Crawford and Howell's, 1998, t-test as explained above), yielded a significant correlation, Phi $= 0.44$, $p = 0.003$, and the point biserial testing the correlation between paraphrasing of object relatives and whether or not hearing aids were fitted by 1 year of age also yielded a significant correlation, $r = 0.42$, $p = 0.001$. Namely, hearing devices (hearing aids or a cochlear implant) fitted by the age of 1 year gave the children a chance of having normal comprehension of relative clauses (as measured by normal paraphrasing of object relatives in our task).

In contrast, **depth of hearing loss** did not correlate with syntactic performance. There were children with profound hearing loss in the normal performance group (12 of the 22 children in this group), and there were children with only medium to severe loss in the syntactically-impaired subgroups (12 of the 26 children in these groups). A Point Biserial test for the correlation between the depth of hearing loss in dB (without hearing aid, measured in the better ear), and the performance of the participant in the object relative paraphrasing task (significantly below the control group or not) showed no significant relation, $r = -0.05$, $p = 0.74$.

The **type of hearing device** the child used also did not correlate with syntactic performance. There were 10 children with hearing aids and 12 with cochlear implants in the normal performance group, and 12 children with hearing aids and 13 with cochlear

[8]One interesting point relates to the performance of the subgroup with the Wh-movement deficit in subject Wh-movement. The table reveals that a few of the Wh-movement impaired individuals also showed (slight) difficulties in production even of subject dependencies. This opens a further question of whether the movement-impairment variety of syntactic impairment can also be sub-divided into children with a deficit only in wh-dependencies in which one lexically-restricted NP crosses another, and children who have a problem with any type of Wh-movement. (These children may still show good performance in comprehension tasks such as sentence-picture matching with subject questions and subject relatives, where they may use an agent-first strategy, a strategy that cannot be employed in production.)

implants in the syntactically-impaired subgroups (9 with cochlear implants, 5 with hearing aids in the CP-impaired group, and 4 with cochlear implants, 7 with hearing aids in the Wh-movement impaired group). Phi Coefficient of Association that was calculated for the type of hearing aid (cochlear implant/hearing aid) and performance in the object relative paraphrasing (significantly or below the control group or not), also yielded no relation between the type of hearing aid and syntactic comprehension, Phi $= -0.08$, Fisher exact $p = 0.77$.

Finally, **Age** at the time of testing within the age group we tested also showed no correlation with performance: the three groups were of the same age ranges, and the point biserial correlation of age in month with performance was not significant: $r = 0.12$, $p = 0.39$.

DISCUSSION

The main questions of this study related to the nature of the syntactic deficit in hearing impairment, a question that took an interesting turn once we analyzed the individual profile of the participants rather than the group's, and to the relation between the syntactic impairment and reading in this population.

ON THE RELATION BETWEEN SYNTACTIC IMPAIRMENT AND READING

This study examined various aspects of the relation between syntactic impairment and reading. The results indicated that hearing impaired children have difficulty understanding sentences derived by Wh-movement, and specifically, object relative clauses, not only when they hear these sentences, but also when the sentences are presented to them in writing, for an unlimited time. The ability of the group of hearing impaired children to interpret the written relative clauses, as reflected in their paraphrases, was significantly worse than that of hearing children. An individual level analysis indicated that 26 of the 48 hearing impaired participants performed significantly worse than the control group in interpreting the relative clause sentences.

This study thus sheds light on reading comprehension in the hearing impaired population. It indicates that individuals with hearing impairment have considerable difficulties in reading comprehension already at the sentence level, and this difficulty is clearly linked to their syntactic impairment, as their paraphrases of the simple control sentences were fine. Some of the hearing impaired participants even showed impaired reading aloud of object relatives, a difficulty that was linked to their syntactic impairment, rather than to a reading impairment. Again, this is supported by their good reading of matched simple sentences. These results open a window to the frequently reported difficulty of hearing impaired children in reading and in written text comprehension. They suggest that the text comprehension difficulties can result from a syntactic impairment.

Furthermore, this study showed that even problems in reading aloud of children with hearing impairment can be ascribed to their syntactic deficit, as the correct reading aloud sometimes requires the correct parsing of the syntactic structure of the sentence that they read. Thus, the syntactic impairment might cause not only difficulties in reading comprehension but also errors in reading decoding, depending on the syntactic structure of the target sentence.

ON THE NATURE OF THE SYNTACTIC IMPAIRMENT IN HEARING IMPAIRED CHILDREN

The major mission of this study was to explore the nature of the deficit in relative clauses in hearing impairment, and specifically to examine whether it is related to syntactic structure or to establishing a chain and assignment of thematic roles to a moved element.

In the test we used, the participants were asked to read aloud an object relative clause in which a homograph was placed after the gap position. This enabled us to see whether the participant was able to construct the syntactic structure of the relative clause and to postulate a trace in the required position. In general, the correct oral reading of heterophonic homographs depends on the semantic content of the sentence and on the correct analysis of its syntactic structure. Because we controlled for the semantic content in our sentences (both meanings of the homograph matched the semantics of the sentence), the oral reading of the homographs provided a sensitive marker for whether or not each participant was able to analyze the syntactic structure of the object relative correctly and to postulate the trace in the correct place.

One of the most striking findings of this study was that it exposed individual differences within the group of the hearing impaired participants who had difficulties in the comprehension of object relatives. Whereas some of them read the homograph correctly but failed to interpret the sentence, other participants could not even read the homograph. This principled variability could only be exposed when the individual performance was examined. It opens the window to important change in our understanding of the syntactic deficits underlying the comprehension and production difficulties in children with hearing impairment, as it exposes two different types of syntactic profiles.

The performance of the hearing-impaired participants who failed to understand the object relatives (namely, failed to paraphrase them) but read the homographs in the relative clauses correctly indicates that their syntactic structure was unimpaired. They could construct the syntactic tree including the CP node correctly, and could represent the embedding in the structure. Therefore, they could identify that there was a moved element (a filler) that they needed to find its original position (gap/trace). However, they could not link the original position (the gap) to the correct moved element and hence, could not reconstruct theta mapping and could not understand the role of the moved element in the embedded sentence. In processing terms (Nicol and Swinney, 1989), one may conceptualize their problem in the following way: they identified the gap position but could not reactivate the correct antecedent there. Therefore, as it were, they could not "undo" the movement operation (across an intervening NP), by reactivating the moved NP at the site of the gap. If one assumes that in online comprehension of an object relative the moved NP is kept in a syntactic STM store until it receives its thematic role, this difficulty can be a difficulty in selecting the NP to be reactivated between two similar NPs in the short-term syntactic store.

The children who failed to read the homograph, we suggest, had not identified the gap position from which the element has moved. Therefore these children not only fail to assign the thematic role to the moved element, they even cannot build the

correct syntactic structure and do not assume the trace. We suggest that the deficit of this subgroup of hearing impaired children lies in the inability to construct the syntactic tree up to its highest nodes. Because they cannot construct the CP, the highest node of the tree, which hosts embedding and the moved element in object relatives, they do not know that they should expect a gap, and hence they completely fail to parse the sentence, and do not assume a trace. Support for their deficit in constructing the CP level came from their performance in the reading and paraphrasing task, as well from their performance in other tasks that involve the CP. In the reading and paraphrasing task these children often omitted and ignored the embedding marker or replaced it with another word. In the other tasks they showed, differently from the other hearing impaired children, a significant difficulty in the production of embedding markers, in elicited production and sentence repetition tasks. They also showed difficulty in the repetition of sentences with verb movement to CP, which the other hearing impaired children repeated correctly. Finally, they also showed a special pattern with respect to subject Wh-dependencies: Other hearing impaired children, such as the ones who read well but could not understand object relatives, typically find it difficult to produce object relatives and object questions, in which one NP is moved over another NP, but produce normally subject relatives and subject questions, in which the movement does not cross another NP. The 15 children with the CP deficit showed significant difficulty not only in the production of object relatives and object questions, but also in the production of subject relatives and subject questions. This indicates again that their difficulty was in the construction of the sentence with the CP node, which is required for both subject and object relative clauses and Wh questions (and not only in Wh questions in which one lexically-restricted DP crosses another).

Thus, we identify two profiles of syntactic deficit in children with hearing impairment: difficulty in creating the link between the moved NP and it original position (the trace), resulting in impaired understanding of the thematic role of the moved NP, and a deficit in building the syntactic structure up to its highest nodes. Researchers who tried to characterize the syntactic deficit of hearing impaired children suggested two different sources for the deficit: Friedmann and Szterman (2006, 2011) argue for a deficit in identifying the thematic roles in sentences in which a NP moved across another one, with good syntactic structure building. De Villiers et al. (1994), on the other hand, suggested that the deficit lies in constructing the high nodes of the syntactic tree. The current study shows that these researchers were both wrong and both right. It is incorrect that all hearing impaired children have intact syntactic structure (contrary to what Friedmann and Szterman, 2006, 2011 suggested), but it is also incorrect that all hearing-impaired children have a CP impairment (contrary to what De Villiers et al., 1994 proposed). Each of these characterizations, however, is correct about a different subgroup of hearing impaired children.

Looking at the background factors, one measure was clearly correlated with syntactic performance: the age at which hearing aids (or a cochlear implant) were fitted. Children who had their hearing devices up to 1 year of age showed significantly better syntactic ability than those who received hearing aids or a cochlear implant when they were older than 1 year. The type of hearing device (hearing aid or cochlear implant), depth of hearing loss (medium, severe, or deep), and age did not predict the syntactic performance. We could not find a background factor that determined, within the syntactically impaired children, who will show the movement deficit and who will show the CP deficit. The crucial effect that the age at which hearing aids are fitted has on later development of syntax points to the first year of life as a critical period for first language acquisition. Language input during the first year of life seems to be crucial for the development of normal syntactic abilities. This conclusion was also reached in earlier studies on children with hearing impairment (Szterman and Friedmann, 2003; Friedmann and Szterman, 2006). Other studies, which have tested language in general but not syntax specifically, also identified the age of identification of the hearing loss and age of initiation into intervention services as the most important predictor for normal language development (Apuzzo and Yoshinaga-Itano, 1995; Yoshinaga-Itano and Apuzzo, 1998a,b; Calderon and Naidu, 2000; Moeller, 2000; Mayberry et al., 2002; Mayberry and Lock, 2003; Yoshinaga-Itano, 2003). Evidence that further supports the importance of the first year of life in the normal development of syntactic abilities comes from a different population. Fattal et al. (2011, 2013) reported that children who did not receive thiamine, a vitamin necessary for brain development, during the first year of life showed severe syntactic difficulties when they were 5 and 9 year olds.

Returning to the two profiles of syntactic impairment, studies with other syntactically-impaired populations also show the two different sources for difficulties with relative clauses in two different populations. Children with Syntactic Specific Language Impairment (syntactic SLI) typically show a deficit that is best described as a deficit in Wh-movement, namely, in the assignment of thematic roles to an NP that moved across another NP (Friedmann and Novogrodsky, 2007, 2011; Friedmann et al., in press). Friedmann and Novogrodsky (2007) investigated this difficulty of Hebrew-speaking children with SySLI using the same task we used here. They found that the children with SySLI read the homographs well, but failed to paraphrase the object relatives. The performance in this task was interpreted as indicating a deficit in movement. Namely, the children with SLI could not activate the correct NP at the trace position. When this happened, the SySLI participants failed to assign the correct thematic role to the moved element, and this led to various paraphrases in which the thematic roles were incorrectly assigned to the arguments. This interpretation is supported by a study by Marinis and van der Lely (2004), who used cross-modal lexical priming and found that English-speaking children with SLI were unable to reactivate the antecedent at the Wh-trace position. That is, even if they know where to place an empty category, they do not know to which phrase to link it, and hence they cannot assign the thematic role correctly.

Another population with a syntactic impairment is that of individuals with agrammatic aphasia. Individuals with Broca's agrammatic aphasia show significant difficulties in the comprehension of sentences derived by syntactic movement that result in a non-canonical order of the arguments such as object relative clauses, object Wh questions, and topicalization structures (Zurif

and Caramazza, 1976; Schwartz et al., 1987; Grodzinsky, 1989, 2000; Grodzinsky et al., 1999; Friedmann and Shapiro, 2003). Friedmann et al. (2006) explored the nature of this deficit using the test we used in the current study (a longer version of it), with individuals with agrammatism. They found that all the individuals with agrammatism they tested were severely impaired in reading the homographs when they appeared after the trace in the object relative clauses, and consequently failed to paraphrase them. Friedmann et al. (2006) interpreted this pattern as evidence for a difficulty in constructing the CP node, a deficit that has abundant evidence for from other sentence production tasks (Friedmann, 2001, 2006). Thus, it seems that some of the hearing impaired children who show a deficit in the comprehension of object relatives have a deficit that is similar to that of children with syntactic SLI, whereas others show a deficit that resembles that of individuals with agrammatism.

When the results regarding the deficit in comprehension and the two different profiles of impairment are brought together with the impression and report of the clinicians and special education teachers who work with these children, some important clinical implications emerge. Whereas all the clinicians and teachers of the hearing impaired children easily detect the syntactic difficulties of the children who had the syntactic structure deficit, and these children are classified with "severe language impairment", the syntactic difficulties of the children who read the sentences correctly is more elusive. Because these children can produce embedded sentences well, and can even read sentences correctly, it is harder to suspect that they have a syntactic deficit. It is thus crucial to test the comprehension of semantically reversible sentences derived by movement even for hearing impaired children who do not display an obvious syntactic deficit in their speech and reading aloud.

One final point relates to whether the deficit of the hearing impaired children who showed impaired comprehension in this task indeed related to Wh-movement or rather to center embedding, as all the object relatives in this study were center-embedded. The participants in this study were part of a wide study in which they were also tested for the comprehension of other Wh-movement structures, including Wh questions, topicalized OSV sentences, and final-branching object relatives in sentence-picture matching tasks and in questions on written sentences. The results of these other tasks indicated that each of the participants who failed in the paraphrasing of the object relatives in the reading task also showed difficulties in the comprehension of object Wh-movement without center embedding (significantly poorer performance in the comprehension of at least one of the above structures). This indicates that they had a genuine deficit in Wh-movement rather than a deficit related to center embedding.

This study thus shows not only that hearing impaired children have syntactic difficulties, but also that these difficulties can have different faces. These difficulties can result from an impairment in syntactic structure building or from an impairment in chain formation: a deficit in linking a moved NP to its original position and hence, impaired assignment of thematic role assignment to the moved NP. The study further suggests that the reading comprehension and reading aloud difficulties can be tightly related to syntactic difficulties, and shows the importance of paying attention to variability within language-impaired groups (see also Coltheart, in press), by analysis of individual performance.

REFERENCES

Allen, T. E. (1986). "Patterns of academic achievement among hearing impaired students: 1974 and 1983," in *Deaf Children in America,* eds A. N. Schildroth and M. A. Karchmer (San Diego, CA: College Hill press), 161–206.

Apuzzo, M. L., and Yoshinaga-Itano, C. (1995). Early identification of infants' significant hearing loss and the Minnesota Child Development Inventory. *Semin. Hear.* 16, 124–139. doi: 10.1055/s-0028-1083710

Berent, G. P. (1988). "An assessment of syntactic capabilities," in *Language Learning and Deafness*, ed M. Strong (Cambridge: Cambridge University Press), 133–161. doi: 10.1017/CBO9781139524483.008

Berent, P. (1996a). "The acquisition of English syntax by deaf learners," in *Handbook of Second Language Acquisition*, eds W. C. Ritchie and T. K. Bhatia (San Diego, CA: Academic Press), 469–506.

Berent, P. (1996b). Learnability constraints on deaf learners' acquisition of English wh-questions. *J. Speech Hear. Res.* 39, 625–643. doi: 10.1044/jshr.3903.625

Calderon, R., and Naidu, S. (2000). Further support of the benefits of early identification and intervention with children with hearing loss. *Volta Rev.* 100, 53–84.

Chomsky, N. (1981). *Lectures on Government and Binding: the Pisa Lectures*. Holland: Foris.

Chomsky, N. (1986). *Knowledge of Language: its Nature, Origin and Use*. Westport, CT: Praeger.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

Chomsky, N. (2000). "Minimalist inquiries," in *Step by Step: Essays on Minimalist Syntax in Honor of Howard Lasnik*, eds R. Martin, D. Michaels, and J. Uriagereka (Cambridge: MIT Press), 89–155.

Chomsky, N. (2001). *Beyond Explanatory Adequacy*. Cambridge, MA: MIT.

Coltheart, M. (in press). "The classic study of Marshall and Newcombe (1973): modern cognitive neuropsychology grew from this paper," in *Cognitive Psychology: Revisiting the Classic Studies,* eds M. W. Eysenck and D. Groome (Thousand Oaks, CA: Sage Publications).

Crawford, J. R., and Garthwaite, P. H. (2002). Investigation of the single case in neuropsychology: confidence limits on the abnormality of test scores and test score differences. *Neuropsychologia* 40, 1196–1208. doi: 10.1016/S0028-3932(01)00224-X

Crawford, J. R., and Howell, D. C. (1998). Regression equations in clinical neuropsychology: an evaluation of statistical methods for comparing predicted and observed scores. *J. Clin. Exp. Neuropsychol.* 20, 755–762. doi: 10.1076/jcen.20.5.755.1132

De Villiers, J., De Villiers, P., and Hoban, E. (1994). "The central problem of functional categories in English syntax of oral deaf children," in *Constraints on Language Acquisition: Studies of Atypical Children,* ed H. Tager Flusberg (Hillsdale, NJ: Erlbaum), 9–47.

De Villiers, P. A. (1988). Assessing English syntax in hearing-impaired children: elicited production in pragmatically motivated situations. *J. Acad. Rehabil. Audiol.* 21(Mono Suppl.), 41–71.

Doron, E., and Meir, I. (2013). "Construct state: modern Hebrew,"in *Encyclopedia of Hebrew Language and Linguistics,* ed G. Khan (Leiden: Brill Online), 581–589.

Fattal, I., Friedmann, N., and Fattal-Valevski, A. (2011). The crucial role of thiamine in the development of syntax and lexical retrieval: a study of infantile thiamine deficiency. *Brain* 134, 1720–1739. doi: 10.1093/brain/awr068

Fattal, I., Friedmann, N., and Fattal-Valevski, A. (2013). "Language abilities of 8–9 year old children who suffered thiamine-deficiency in infancy," in *Presented in the 49th Annual Conference of the Israeli Speech Hearing and Language Association* (Tel Aviv).

Friedmann, N. (2000). *PETEL: a Sentence Repetition Test*. Tel Aviv: Tel Aviv University.

Friedmann, N. (2001). Agrammatism and the psychological reality of the syntactic tree. *J. Psycholinguist. Res.* 30, 71–90. doi: 10.1023/A:1005256224207

Friedmann, N. (2006). "Speech production in Broca's agrammatic aphasia: syntactic tree pruning," in *Broca's Region,* eds Y. Grodzinsky and K. Amunts (New York, NY: Oxford University Press), 63–82.

Friedmann, N., and Gvion, A. (2003). *TILTAN: A Test Battery for Dyslexias*. Tel Aviv: Tel Aviv University.

Friedmann, N., Gvion, A., and Novogrodsky, R. (2006). "Syntactic movement in agrammatism and S-SLI: two different impairments," in *Language Acquisition and Development,* eds A. Belletti, E. Bennati, C. Chesi, E. Di Domenico, and I. Ferrari (Cambridge, UK: Cambridge Scholars Press/CSP), 197–210.

Friedmann, N., and Haddad-Hanna, M. (2014). The comprehension of sentences derived by syntactic movement in Palestinian Arabic-speaking children with hearing impairment. *Appl. Psycholinguist*. 35, 473–513. doi: 10.1017/S0142716412000483

Friedmann, N., and Lukov, L. (2008). Developmental surface dyslexias. *Cortex* 44, 1146–1160. doi: 10.1016/j.cortex.2007.09.005

Friedmann, N., and Novogrodsky, R. (2007). Is the movement deficit in syntactic SLI related to traces or to thematic role transfer? *Brain Lang.* 101, 50–63. doi: 10.1016/j.bandl.2006.09.006

Friedmann, N., and Novogrodsky, R. (2011). Which questions are most difficult to understand? The comprehension of Wh questions in three subtypes of SLI. *Lingua* 121, 367–382. doi: 10.1016/j.lingua.2010.10.004

Friedmann, N., and Shapiro, L. P. (2003). Agrammatic comprehension of simple active sentences with moved constituents: Hebrew OSV and OVS structures. *J. Speech Lang. Hear. Res.* 46, 288–297. doi: 10.1044/1092-4388(2003/023)

Friedmann, N., and Szterman, R. (2006). Syntactic movement in orally-trained children with hearing impairment. *J. Deaf Stud. Deaf Educ.* 11, 56–75. doi: 10.1093/deafed/enj002

Friedmann, N., and Szterman, R. (2011). The comprehension and production of Wh questions in children with hearing impairment. *J. Deaf Stud. Deaf Educ.* 16, 212–235. doi: 10.1093/deafed/enq052

Friedmann, N., Szterman, R., and Haddad-Hanna, M. (2010). "The comprehension of relative clauses and Wh questions in Hebrew and Palestinian Arabic hearing impairment," in *Language Acquisition and Development: Generative Approaches to Language Acquisition 2009,* eds A. Castro, J. Costa, M. Lobo, and F. Pratas (Cambridge, UK: Cambridge Scholars Press/CSP), 157–169.

Friedmann, N., Yachini, M., and Szterman, R. (in press). "Relatively easy relatives: children with syntactic SLI avoid intervention," in *Structures, Strategies and Beyond. Studies in Honour of Adriana Belletti*, eds E. Di Domenico, C. Hamann, and S. Matteini (Amsterdam: John Benjamins; Linguistik Aktuell series).

Grodzinsky, Y. (1989). Agrammatic comprehension of relative clauses. *Brain Lang.* 37, 480–499. doi: 10.1016/0093-934X(89)90031-X

Grodzinsky, Y. (2000). The neurology of syntax: language use without Broca's erea. *Behav. Brain Sci*. 23, 1–71. doi: 10.1017/S0140525X00002399

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.

Haddad-Hanna, M., and Friedmann, N. (2009). The comprehension of syntactic structures by Palestinian Arabic-speaking individuals with hearing impairment. *Lang. Brain* 9, 79–104 (in Arabic).

Haddad-Hanna, M., and Friedmann, N. (2014). "The comprehension and production of sentences with syntactic movement in Palestinian Arabic-speaking individuals with hearing impairment," in *Rehabilitation and Education of Children and Adults Hard of Hearing and Deaf: Theoretical and Implementation Aspects,* eds T. Most and D. Ringwald-Frimermen (Tel Aviv: Mofet institute) (in Hebrew), 295–351.

Kayne, R. (1994). *The Antisymmetry of Syntax*. Cambridge, MA: MIT Press.

Kesselman, A., Yachini, M., and Friedmann, N. (2013). "Dyslexia and ısyntactic SLI: together or separate?" in *Presented at the 7th Annual Conference of the Israeli Association for Literacy and Language* (Kiryat Ono).

Luckner, J. L., and Handley, C. M. (2008). A Summary of the reading comprehension research undertaken with students who are deaf or hard of hearing. *Am. Ann. Deaf* 153, 6–36. doi: 10.1353/aad.0.0006

Marinis, T., and Van der Lely, H. K. J. (2004). "The underlying representation of Wh-questions in subgroups of children with SLI: evidence from on-line sentence processing," in *Presented at the 29th Boston University Conference on Language Development* (Boston, MA).

Mayberry, R. I., Chen, J. K., Witcher, P., and Klein, D. (2011). Age of acquisition effects on the functional organization of language in the adult brain. *Brain Lang.* 119, 16–29. doi: 10.1016/j.bandl.2011.05.007

Mayberry, R. I., and Lock, E. (2003). Age constraints on first versus second language acquisition: evidence for linguistic plasticity and epigenesis. *Brain Lang.* 87, 369–383 doi: 10.1016/S0093-934X(03)00137-8

Mayberry, R., Lock, E., and Kazmi, H. (2002). Development: linguistic ability and early language exposure. *Nature* 417:38. doi: 10.1038/417038a

Moeller, M. P. (2000). Early intervention and language development in children who are deaf and hard of hearing. *Pediatrics* 106:e43. doi: 10.1542/peds.106.3.e43

Moeller, M., Tomblin, B., Yoshinaga-Itano, C., McDonald Connor, C., and Jerger, S. (2006). Current state of knowledge: language and literacy of children with hearing impairment. *Ear Hear.* 28, 740–753. doi: 10.1097/AUD.0b013e318157f07f

Moog, J., and Geers, A. (1985). EPIC: a program to accelerate academic progress in profoundly hearing-impaired children. *Volta Rev*. 87, 259–277.

Musselman, C. (2000). How do children who can't hear learn to read an alphabetic script? A review of the literature on reading and deafness. *J. Deaf Stud. Deaf Educ.* 5, 9–31. doi: 10.1093/deafed/5.1.9

Nave, M., Szterman, R., and Friedmann, N. (2009). Comprehension and production of Wh questions by Hebrew-speaking children with hearing impairment: another evidence for the difficulty in syntactic movement. *Lang. Brain* 9, 1–29 (in Hebrew).

Nicol, J., and Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *J. Psycholinguist. Res.* 18, 5–19. doi: 10.1007/BF01069043

Novogrodsky, R., and Friedmann, N. (2006). The production of relative clauses in SLI: a window to the nature of the impairment. *Adv. Speech Lang. Pathol.* 8, 364–375. doi: 10.1080/14417040600919496

Onifer, W., and Swinney, D. (1981). Accessing lexical ambiguities during sentence comprehension: effects of frequency of meaning and context bias. *Mem. Cogn.* 9, 225–236. doi: 10.3758/BF03196957

Pollock, J. Y. (1989). Verb movement, universal grammar and the structure of IP. *Linguist. Inq.* 20, 365–424.

Quigley, S. P., Smith, N. L., and Wilbur, R. B. (1974a). Comprehension of relativized sentences by deaf students. *J. Speech Hear. Res.* 17, 325–341. doi: 10.1044/jshr.1703.325

Quigley, S. P., Wilbur, R. B., and Montanelli, D. S. (1974b). Questions formation in the language of deaf students. *J. Speech Hear. Res.* 17, 699–713. doi: 10.1044/jshr.1704.699

Rizzi, L. (1990). *Relativized Minimality*. Cambridge, MA: MIT Press.

Rizzi, L. (1997). "The fine structure of the left periphery," in *Elements of Grammar: Handbook in Generative Syntax*, ed L. Haegeman (Dordrecht: Kluwer), 281–337.

Sauerland, U. (2000). "Two structures for English restrictive relative clauses," in *Proceedings of the Nanzan GLOW,* eds M. Saito, Y. Abe, H. Aoyagi, J. Arimoto, K. Murasugi, and T. Suzuki (Nagoya: Nanzan University), 351–366.

Schwartz, M. F., Linebarger, M. C., Saffran, E. M., and Pate, D. S. (1987). Syntactic transparency and sentence interpretation in aphasia. *Lang. Cogn. Processes* 2, 85–113. doi: 10.1080/01690968708406352

Szterman, R., and Friedmann, N. (2003). The deficit in comprehension of movement-derived sentences in children with hearing impairment. *Lir'ot et Hakolot* 2, 20–29 (in Hebrew).

Szterman, R., and Friedmann, N. (2007). How do children with hearing impairment produce relative clauses? *Isr. J. Lang. Speech Hear. Disord.* 28, 58–71 (in Hebrew).

Szterman, R., and Friedmann, N. (2011). "The comprehension of verb movement in hearing impaired children and typically developing children," in *Presented at the 10th IMAM Conference, Language and Brain Lab* (Tel Aviv: Tel Aviv University).

Szterman, R., and Friedmann, N. (2014). "The syntactic abilities of children with hearing impairment in school ages and its influence on reading comprehension," in *Rehabilitation and Education and of Children and Adults Hard of Hearing and Deaf: Theoretical and Implementational Aspects,* eds T. Most and D. Ringwald-Frimermen (Tel Aviv: Mofet institute) (in Hebrew), 239–294.

Traxler, C. (2000). The stanford achievement test, 9th edition: national norming and performance standards for deaf and hard-of-hearing students. *J. Deaf Stud. Deaf Educ.* 5, 337–345. doi: 10.1093/deafed/5.4.337

Trybus, R., and Karchmer, M. (1977). School achievement scores of hearing impaired children: national data on achievement status and growth patterns. *Am. Ann. Deaf* 122, 62–69.

Vergnaud, J. R. (1974). *French Relative Clauses*. Ph.D. dissertation, MIT.

Volpato, F., and Adani, F. (2009). "The subject/object relative clause asymmetry in Italian hearing-impaired children: evidence from a comprehension task," in *Proceedings of the XXXV Incontro di Grammatica Generativa*, ed V. Moscati (Siena).

Yoshinaga-Itano, C. (2003). From screening to early identification and intervention: discovering predictors to successful outcomes for children with significant hearing loss. *J. Deaf Stud. Deaf Educ.* 8, 11–30. doi: 10.1093/deafed/8.1.11

Yoshinaga-Itano, C., and Apuzzo, M. L. (1998a). Identification of hearing loss after age 18 months is not early enough. *Am. Ann. Deaf* 143, 380–387. doi: 10.1353/aad.2012.0151

Yoshinaga-Itano, C., and Apuzzo, M. L. (1998b). The development of deaf and hard of hearing children identified early through the high-risk registry. *Am. Ann. Deaf* 143, 416–424. doi: 10.1353/aad.2012.0118

Zurif, E., and Caramazza, A. (1976). "Psycholinguistic structures in aphasia: studies in syntax and semantics," in *Studies in Neurolinguistics*, Vol. 1, eds H. Whitaker and H. A. Whitaker (New York, NY: Academic Press), 260–292.

Zurif, E., Swinney, D., Prather, P., Solomon, J., and Bushell, C. (1993). An on-line analysis of syntactic processing in Broca's and Wernicke's aphasia. *Brain Lang.* 45, 448–464. doi: 10.1006/brln.1993.1054

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.

Received: 19 January 2014; accepted: 09 October 2014; published online: 07 November 2014.

Citation: Szterman R and Friedmann N (2014) Relative clause reading in hearing impairment: different profiles of syntactic impairment. Front. Psychol. ***5****:1229. doi: 10.3389/fpsyg.2014.01229*

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

frontiers in
PSYCHOLOGY

MINI REVIEW ARTICLE
published: 29 July 2014
doi: 10.3389/fpsyg.2014.00833

Colors, colored overlays, and reading skills

Arcangelo Uccula[1]*, Mauro Enna[1] and Claudio Mulatti[2]

[1] *Dipartimento di Storia, Scienze dell'Uomo e della Formazione, Università degli Studi di Sassari, Sassari, Italy*
[2] *Dipartimento di Psicologia dello Sviluppo e della Socializzazione, Università degli Studi di Padova, Padova, Italy*

Edited by:
Simona Amenta, University of Milano-Bicocca, Italy

Reviewed by:
Olaf Hauk, MRC Cognition and Brain Sciences Unit, UK
Joana Acha, Basque Centre on Cognition, Brain and Language, Spain
Laura Danelli, Università degli studi di Milano-Bicocca, Italy

****Correspondence:***
Arcangelo Uccula, Dipartimento di Storia, Scienze dell'Uomo e della Formazione, Università degli Studi di Sassari, Piazza Conte di Moriana 8, Sassari 07100, Italy
e-mail: uccula@uniss.it

In this article, we are concerned with the role of colors in reading written texts. It has been argued that colored overlays applied above written texts positively influence both reading fluency and reading speed. These effects would be particularly evident for those individuals affected by the so called Meares-Irlen syndrome, i.e., who experience eyestrain and/or visual distortions – e.g., color, shape, or movement illusions – while reading. This condition would interest the 12–14% of the general population and up to the 46% of the dyslexic population. Thus, colored overlays have been largely employed as a remedy for some aspects of the difficulties in reading experienced by dyslexic individuals, as fluency and speed. Despite the wide use of colored overlays, how they exert their effects has not been made clear yet. Also, according to some researchers, the results supporting the efficacy of colored overlays as a tool for helping readers are at least controversial. Furthermore, the very nature of the Meares-Irlen syndrome has been questioned. Here we provide a concise, critical review of the literature.

Keywords: color, overlays, meares-irlen syndrome, reading, dyslexia

INTRODUCTION

The role of colors in reading has a few decades of history, dating back to 1958, when Jansky (1958) reported the case of a student with a reading deficit who was unable to recognize words printed on a white paper but was able to recognize words printed on a yellow paper. Although the theoretical debate on the causes of reading difficulties and dyslexia has given a primary role to the "phonological hypothesis" – since the efficiency of the processes of phonological processing is among the best predictors of reading skill acquisition (Wagner and Torgesen, 1987; Snowling et al., 2000) – the role of visual and perceptual skills has gained attention (e.g., Watson et al., 2003). One of the reasons that brought the role of visual and perceptual skills in reading to attention was the observation that some dyslexic individuals are affected by a perceptual dysfunction, called Scotopic Sensitivity Syndrome and also known as Meares-Irlen Syndrome and Visual Stress (MISViS; Evans, 1997).

In this article, we provide a brief, concise review of the literature on colored overlays as a remedy for visual stress in reading. To forecast the conclusions, the conception of visual stress as an independent reading deficit is controversial, whereas the research on the colored overlays is yet inconclusive since evidence both in favor of and contrary to their efficacy as a remedy has been provided.

VISUAL STRESS AND READING

The term "visual stress" refers to the inability to see comfortably and without distortion (Wilkins et al., 1984). With "visual stress" Wilkins refers to the condition caused by the features of the visual stimulus, and that therefore is of sensorial origin, and not to the visual stress generated by movements of the eyes, by visual accommodation or by binocular convergence. Symptoms of visual stress are visual fatigue, perceived excessive luminosity, and several kinds of perceptual distortion such as blurring, fading, or flickering of the visual stimulus. According to Irlen (1997), this condition would interest approximately the 12–14% of the population and about the 46% of individuals with a diagnosis of dyslexia (and/or alternative learning difficulties). A more recent study (Kriss and Evans, 2005) suggests that visual stress affects about the 37.5% of children with dyslexia and about the 25% of non-dyslexic children. The frequencies of the symptoms would be: blurring (24%), duplication (16%), jump (12%), format switch (6%), and fading (3.5%) of the visual stimulus (Kriss and Evans, 2005).

According to Meares (1980), the factors that contribute most to the reading difficulties in children originates in the perceptual instability of the visual input due to the organization of the figure with respect to the background of the black ink writing on a white paper, which is typical in printed books. The idea, therefore, is that for some individuals the reflex of the black ink on a white paper makes reading difficult.

COLORED OVERLAYS

The idea here is that if visual stress is the result of the relation between the visual features of black ink writing on white paper, then changing this relation might result in a reduction of the symptoms associated with visual stress (c.f. Wilkins, 2003; Irlen, 2010). A way to alter the relation between the visual features of the written text and the background is to place on the text a colored sheet of transparent plastic (colored overlay). Scott et al. (2002; see also Kruk et al., 2008) had shown that whereas bad readers show – after about 10 min of reading black writing on a white paper – the typical symptoms of visual stress, they do not show visual stress' symptoms when reading the texts with the same characteristics through a colored overlay.

Thus, the implications of the colored overlays method is that if visual stress impairs reading acquisition, then the use of colored overlays might improve both reading and reading acquisition (Irlen, 2010).

DYSLEXIA AND COLORED OVERLAYS

According to Evans et al. (1999) colored filters determines benefit in about 80% of individuals using them. The adoption of colored overlays/filters in schools is incremented given that the visual stress syndrome – which symptoms they are supposed to alleviate – is often observed in dyslexic students (Irlen, 1991; Singleton and Trotter, 2005; Singleton and Henderson, 2007), and it is in schools that students are usually diagnosed as dyslexics. The estimation of visual stress is, in fact, often included in tests aimed at assessing reading skills and dyslexia (Nichols et al., 2009), and the colored overlays are often used as a remedy for the visual stress symptoms co-occurring with dyslexia. However, several studies have shown that dyslexia and visual stress are independent conditions. Originally, in fact, visual stress was considered as a subset of dyslexia, whereas more recently it has been argued that the visual stress syndrome is independent from dyslexia (Kriss and Evans, 2005; Kruk et al., 2008). Indeed Kriss and Evans (2005) noted that the prevalence of visual stress in dyslexic individuals is of only 10% higher than in the non-dyslexic individuals: from this the authors conclude that dyslexia and visual stress are two independent conditions which sometimes coexist within the same individual.

Although dyslexia and visual stress seem independent syndromes, it is often the case that significantly large sub-groups of dyslexics do have deficits in visual processing (Watson and Willows, 1995), and when dyslexia is associated with a visual-perceptual deficit, reading difficulties worsen (Wilkins et al., 2001). In fact, it has been shown that when dyslexic children can read through a self-chosen colored overlay, they reading speed increases of about a 25% (Wilkins, 2002): moreover, although it seems that even non-dyslexic children benefit from the use of colored overlays, the benefit resulting from the use of colored overlays by dyslexic children is higher than that observed with non-dyslexic children (Singleton and Henderson, 2007). With respect to adults, it seems that only individuals with dyslexia and visual stress syndrome benefit from the use of colored overlays when compared with dyslexics without visual stress, non-dyslexics with visual stress, and non-dyslexics without visual stress.

Singleton and Trotter (2005) classified a sample of dyslexic and non-dyslexic individuals as a function of whether they experienced high or low intensities of visual stress, and observed that only the dyslexics individuals experiencing visual stress of high intensity benefitted from colored overlays. From this, the authors concluded that dyslexia and visual stress are related: they argued that if the two conditions were independent, as proposed by Wilkins, all individuals experiencing intense visual stress should have benefitted from colored overlays, regardless the concurrent presence of dyslexia. Noteworthy, the argument of Singleton and Trotter assumes that colored overlays were always beneficial for visual stress, when in presence of visual stress, and since colored overlay are not beneficial for not dyslexics individuals with intense visual stress, visual stress and dyslexia are inter-dependent. But of course, one could argue here that it is the efficacy of colored overlay that depends on the coexistence of the two conditions, regardless of whether or not the two conditions are dependent.

Thus, there exist two views. According to one view, visual stress and dyslexia are independent conditions. According to the other view, visual stress and dyslexia are dependent conditions.

HOW DOES COLOR HELP READING (IF IT DOES)?

Despite the many studies aimed at investigating the role of colors in reading – also altering the features of the letters (Pinna et al., 2010) – and that the colored overlay are widely used, the mechanisms at the basis of the relation between reading and color have not been properly understood. Possibly, one of the reason for this lack of explanations is that the very nature of the visual stress syndrome and of its role in reading has been questioned, and therefore the entire enterprise might just be a false trail.

A recent account of the causes of visual stress posits that a strong sensorial stimulation – as a dense written text – might lead to a reduction of the efficiency of the inhibitory mechanisms in the visual cortex, thus resulting in an excessive excitation of the cortical neurons, and this would cause illusions and distortions (Huang et al., 2003). This hypothesis implies that some individuals are affected by a sort of cortical hypersensitivity so that their visual cortex would overreact to intense visual stimulations thus determining the symptoms associated with visual stress, as fatigue and migraine. Building on this ground, Wilkins and Evans (2010) proposed that the colored overlays are effective because they distribute this excessive excitation and thus mitigate the symptoms of visual stress, thus improving written text processing and reading. Although this account lacks of strong empirical evidence (Henderson et al., 2013), a recent neuroimaging study by Chouinard et al. (2012) provides some initial evidence showing cortical over-excitability in presence of visual stress syndrome.

This view of the basis of visual stress is congruent with early studies (Wilkins et al., 1994; Robinson and Foreman, 1999) showing that the color of the colored overlay is specific for each individual, that is whit the fact that each reader benefits from the use of colored overlays only if the color of the overlay is a specific color.

Some of the symptoms of visual stress as blurring and illusory migrations of letters are similar to those reported in presence of magnocellular dysfunctions (Stein and Walsh, 1997). A dysfunction of the magnocellular pathway would produce long lasting, anomalous visual traces which would interfere – by masking – with the visual processing of the stimulation thus causing blurring and distortions. The empirical evidence here is once more inconsistent (Skoyles and Skottun, 2009).

Wilkins (2003) argues that the hypothesis of a magnocellular dysfunction at the basis of visual stress might account for the individual differences in the use of colors – this because it has been shown that each individual benefits from the use of a given, specific color, not from any possible color. This last proposal lacks of empirical evidence.

According to some authors, the candidate brain structure for understanding the relation between colored overlays and reading is the magnocellular system (Chase et al., 2003). In fact, it has been shown that reading is compromised within a red light environment compared to a green light environment, this because the red light inhibits the activity of the magnocellular system (Chase et al., 2003). Similarly, Ray et al. (2005) have shown that yellow filters – by reducing the blue components of the light which inhibit the

activity of the magnocellular system – increase the ability of reading in dyslexic populations (however, this has not been replicated, see: Palomo-Álvarez and Puell, 2013). Although these findings are consistent with the idea that reading proficiency benefits from the use of colored filters, they are inconsistent with early findings on colored overlays, since early findings show that each individual benefits from the use of a particular, given color, whereas these latter findings suggest that a particular color – e.g., yellow – should work for any reader.

LATEST DEVELOPMENTS ON COLORED OVERLAYS: DO THEY WORK OR NOT?

In recent studies, serious methodological limits in the works supporting the use of colored overlays have been pointed out.

One of the main methodological issues has to do with the definition and the diagnosis of visual stress and originates in the way visual stress is assessed. Some authors diagnose or not visual stress as a function of how participants responds to treatments based on colored overlays (Kriss and Evans, 2005). Others instead stress the symptoms of visual stress as visual distortions in reading (Singleton and Trotter, 2005). It has been noted that in order to use the improvements in reading due to the use of colored overlays as a diagnostic criterion, the symptoms should be univocally attributable to the Meares-Irlen syndrome, which is not necessarily the case (Kruk et al., 2008). In addition, some have considered a 20% increase in reading speed with the use of colored overlays as a threshold for the diagnosis of visual stress (Minwook et al., 2014), others used a 5% increase in reading speed as criterion. Of course, the prevalence of the Meares-Irlen syndrome changes as a function of the threshold used. Wilkins et al. (2001) found that with a threshold of 5% of increase in reading speed due to colored overlays, the 33% of 6–8 years old children suffers of visual stress. With a threshold of 10%, the prevalence falls to 12.5% (Kriss and Evans, 2005), whereas with a threshold of 25%, the prevalence falls to 5% (Wilkins et al., 2001). The prevalence of visual stress increases if the samples are limited to dyslexic individuals, and goes from 47% with a threshold of 5–31% with a threshold of 10%.

Noteworthy, the evaluation of the symptoms is based on subjective reports, and, in the studies of Wilkins and colleagues (e.g., Wilkins et al., 2005), the participants select their favorite colors or combination of colors themselves. From one side, these aspects question the reliability of the diagnosis, as confirmed by low test-retest reliability (Woerz and Maples, 1997). From another side, these specificity and variability in the selection of the color complicate the search for an explanation of why a color is better than another for a given individual, especially assuming visual stress is a unique condition.

Some recent studies failed to find statistically significant effects of colored overlays. Ritchie et al. (2011) had shown that, in the short period, colored overlays do not speed reading up compared to non-colored overlays, whether or not the participants have a diagnosis of visual stress. Ritchie et al. (2012) had shown that – compared to a control condition – not even one year of use of colored overlays results in an increase in reading speed and accuracy. Henderson et al. (2013) had shown that despite the fact that often dyslexic individuals do experience stronger visual stress than controls, neither dyslexics nor controls benefit from the use of colored overlays.

DISCUSSION AND CONCLUSION

The existence itself of the visual stress syndrome is – at least as an independent condition – controversial: symptoms that have been considered mapping into an independent cluster might just be individual-specific aspects of the more wide and articulated dyslexia. In addition, typical visual stress symptoms might be symptoms of dyslexia rather than causes (Olitsky and Nelson, 2003), and thus the attenuation of those symptoms – whatever the technique employed – might have no consequences on the quality of the reading. It has been shown that children with reading difficulties do like to play video games and do play video games for long time: some have argued that if at the basis of their reading difficulties there were perceptual deficits, then they would avoid such high intensity visual activities as video gaming [American Academy of Pediatrics (AAP), 2009]. However, it has been shown that playing action video games improves reading skills of dyslexic children more than traditional reading treatments, possibly because action video games improve attentional abilities (Franceschini et al., 2013). This implies that despite their lower attentional abilities, dyslexic children do like to play video games and, also, obtain benefits from playing video games. Thus, if the visual stress existed, then – analogously – children with visual stress might not only like to play video games, they might also obtain benefits from playing video games.

The idea at the basis of the Meares-Irlen syndrome, whether or not the syndrome exists as an independent collection of symptoms, contributed – by focusing on the early, input processes – to the identification of visual disorders which have been observed to occur in presence of reading difficulties or dyslexia, thus contrasting the dominant view of dyslexia that sees the deficit as due to phonological processing impairments (Ramus, 2014). For example, in a recent, single-case study of a dyslexic children, the authors found visual processing disorders but not phonological disorders (Valdois et al., 2011).

Whether colored overlays help reading or not seems at least controversial: although initial evidence was indeed provided, more recent studies both highlight the methodological issue of previous studies and show that colored overlays do not help reading (Ritchie et al., 2011, Ritchie et al., 2012; Henderson et al., 2013), On the ground of contradictory findings as these, the [American Academy of Pediatrics (AAP), 2009] has claimed that there is not empirical evidence toward the efficacy of colored overlays in reading, reading acquisition, or dyslexia, and did not recommend their use.

However, the participants in the studies of Ritchie et al. (2011); Ritchie et al. (2012) were non-dyslexic children, and in the study of Henderson et al. (2013) they were adults, whereas it has been shown that effects of colored overlays are more easily found with dyslexic children (Singleton and Trotter, 2005; Singleton and Henderson, 2007). Whether or not, at least in some conditions, colored overlay works does not seem to be a settled issue. Thus, although from one side, given these contradictory findings, a precautionary, prudent position – as that of the Academy of Pediatrics – on the use of colored overlay seems desirable, especially in clinical or educative contexts, from another side,

given that some evidence that the colored overlays work exists, concluding that colored overlays proved not worth in allaying reading problems is premature and, possibly, incorrect.

REFERENCES

American Academy of Pediatrics [AAP]. (2009). Joint statement: learning disabilities, dyslexia, and vision. *Pediatrics* 124, 837–844. doi: 10.1542/peds.2009-1445

Chase, C., Ashourzadeh, A., Kelly, C., Monfette, S., and Kinsey, K. (2003). Can the magnocellular pathway read? Evidence from studies of colour. *Vision. Res.* 43, 1211–1222. doi: 10.1016/S0042-6989(03)00085-3

Chouinard, B. D., Zhou, C. I., Hrybouski, S., Kim, E. S., and Cummine, J. (2012). A functional neuroimaging case study of Meares-Irlen syndrome/visual stress (MISViS). *Brain Topogr.* 25, 293–307. doi: 10.1007/s10548-011-0212-z

Evans, B. J. W. (1997). Coloured filters and dyslexia: what's in a name? *Dyslexia Rev.* 9, 18–19.

Evans, B. J. W., Patel, R., Wilkins, A. J., Lightstone, A., Eperjesi, F., Speedwell, L., et al. (1999). A review of the management of 323 consecutive patients seen in a specific learning difficulties clinic. *Ophthal. Physl. Opt.* 19, 454–466. doi: 10.1046/j.1475-1313.1999.00465.x

Franceschini, S., Gori, S., Ruffino, M., Viola, S., Molteni, M., and Facoetti, A. (2013). Action video games make dyslexic children read better. *Curr. Biol.* 23, 462–466. doi: 10.1016/j.cub.2013.01.044

Henderson, L. M., Tsogka, N., and Snowling, M. J. (2013). Questioning the benefits that coloured overlays can have for reading in students with and without dyslexia. *Jorsen* 13, 57–65.

Huang, J., Cooper, T. G., Satana, B., Kaufman, D. I., and Cao, Y. (2003). Visual distortion provoked by a stimulus in migraine associated with hyperneuronal activity. *Headache* 43, 664–671. doi: 10.1046/j.1526-4610.2003.03110.x

Irlen, H. (1991). *Scotopic Sensitivity Syndrome: Screening manual.* Long Beach, CA: Perceptual Development Corporation.

Irlen, H. (1997). Reading problems and Irlen coloured lenses. *Dyslexia Rev. Spring* 4–7.

Irlen, H. (2010). *The Irlen Revolution: A Guide to Changing Your Perception and Your Life.* New York, NY: Square One Publishers.

Jansky, J. (1958). A case of severe dyslexia with aphasic-like symptoms. *Bull. Orton Society* 8, 8–11. doi: 10.1007/BF02657600

Kriss, I., and Evans, B. J. W. (2005). The relationship between dyslexia and Meares–Irlen Syndrome. *J. Res. Read.* 28, 350–364. doi: 10.1111/j.1467-9817.2005.00274.x

Kruk, R., Sumbler, K., and Willows, D. (2008). Visual processing characteristics of children with Meares–Irlen syndrome. *Ophthal. Physl. Opt.* 28, 35–46. doi: 10.1111/j.1475-1313.2007.00532.x

Meares, O. (1980). Figure/background, brightness/contrast, and reading disabilities. *Visible Lang.* 14, 13–29.

Minwook, C., Seung-Hyun, K., Joo-Young, K., and Yoonae, A. C. (2014). Specific visual symptoms and signs of meares-irlen syndrome in korean. *Korean J. Ophthalmol.* 28, 159–163. doi: 10.3341/kjo.2014.28.2.159

Nichols, S. A., McLeod, J. S., Holder, R. L., and McLeod, H. S. T. (2009). Screening for dyslexia, dyspraxia and Meares-Irlen syndrome in Higher Education. *Dyslexia* 15, 42–60. doi: 10.1002/dys.382

Olitsky, S. E., and Nelson, L. B. (2003). Rading disorders in children. *Pediatr. Clin. North. Am.* 50, 213–224. doi: 10.1016/S0031-3955(02)00104-9

Palomo-Álvarez, C., and Puell, M. C. (2013). Effects of wearing yellow spectacles on visual skills, reading speed, and visual symptoms in children with reading difficulties. *Graefes Arch. Clin. Exp. Ophthalmol.* 251, 945–951. doi: 10.1007/s00417-012-2162-x

Pinna, B., Uccula, A., and Tanca, M. (2010). How does the color influence figure and shape formation, grouping, numerousness and reading? The role of chromatic wholeness and fragmentation. *Ophthal. Physl. Opt.* 30, 583–593. doi: 10.1111/j.1475-1313.2010.00743.x

Ramus, F. (2014). Neuroimaging sheds new light on the phonological deficit in dyslexia. *Trends Cogn. Sci.* 18, 274–275. doi: 10.1016/j.tics.2014.01.009

Ray, N. J., Fowler, S., and Stein, J. F. (2005). Yellow filters can improve magnocellular function: motion sensitivity, convergence, and reading. *Ann. Ny. Acad. Sci.* 1039, 283–293. doi: 10.1196/annals.1325.027

Ritchie, S. J., Della Sala, S., and McIntosh R. D. (2011). Irlen colored overlays do not alleviate reading difficulties, *Pediatrics* 128, 932–938. doi: 10.1542/peds.2011-0314

Ritchie, S. J., Della Sala, S., and McIntosh, R. D. (2012). Irlen colored filters in the classroom: a 1-Year Follow-Up. *Mind. Brain Educ.* 6, 74–80. doi: 10.1111/j.1751-228X.2012.01139.x

Robinson, G. L., and Foreman, P. J. (1999). Scotopic sensitivity/Irlen Syndrome and the use of coloured filters: a long-term placebo-controlled and masked study of reading achievement and perception of ability. *Percept. Motor Skill* 88, 35–52. doi: 10.2466/pms.1999.88.1.35

Scott, L., McWhinnie, H., Taylor, L., Stevenson, N., Irons, P., Lewis, E., et al. (2002). Coloured overlays in schools: orthoptic and optometric findings. *Ophthal. Physl. Opt.* 22, 156–165. doi: 10.1046/j.1475-1313.2002.00009.x

Singleton, C., and Henderson, L. M. (2007). Computerized screening for visual stress in children with dyslexia. *Dyslexia* 13, 130–151. doi: 10.1002/dys.329

Singleton, C., and Trotter S. (2005). Visual stress in adults with and without dyslexia. *J. Res. Read.* 28, 365–378. doi: 10.1111/j.1467-9817.2005.00275.x

Skoyles, J. R., and Skottun, B. C. (2009). Conflicting data about dyslexia's cause. *Science* 326, 228–229. doi: 10.1126/science.326_228b

Snowling, M., Bishop, D. V., and Stothard, S. E. (2000). Is preschool language impairment a risk factor for dyslexia in adolescence? *J. Child Psychol. Psychiat.* 41, 587–600. doi: 10.1111/1469-7610.00651

Stein, J., and Walsh, V. (1997). To see but not to read; the magnocellular theory of dyslexia. *Trends Neurosci.* 20, 147–152. doi: 10.1016/S0166-2236(96)01005-3

Valdois, S., Bidet-Ildei, C., Lassus-Sangosse, D., Reilhac, C., N'guyen-Morel, M. A., Guinet, E., et al. (2011). A visual processing but no phonological disorder in a child with mixed dyslexia. *Cortex* 47, 1197–1218. doi: 10.1016/j.cortex.2011.05.011

Wagner, R., and Torgesen, J. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychol. Bull.* 101, 192–212. doi: 10.1037/0033-2909.101.2.192

Watson, C. S., Kidd, G. R., Horner, G., Connell, P. J., Lowther, A., Eddins, D. A., et al. (2003). Sensory, cognitive, and linguistic factors in early academic performance of elementary school children: the Benton-IU project. *J. Learn. Disabil.* 36, 165–197. doi: 10.1177/002221940303600209

Watson, C., and Willows, D. M. (1995). Information-processing patterns in specific reading disability. *J. Learn. Disabil.* 28, 216–231. doi: 10.1177/002221949502800404

Wilkins, A. J. (2002). Coloured overlays and their effects on reading speed: a review. *Ophthal. Physl. Opt.* 22, 448–454. doi: 10.1046/j.1475-1313.2002.00079.x

Wilkins, A. J. (2003). *Reading Through Colour.* Chichester: John Wiley and Sons.

Wilkins, A. J., and Evans, B. J. (2010).Visual stress, its treatment with spectral filters, and its relationship to visually induced motion sickness. *Appl. Ergon.* 41, 509–515. doi: 10.1016/j.apergo.2009.01.011

Wilkins, A. J., Evans, B. J. W., Brown, J., Busby, A., Wingfield, A. E., Jeanes, R., et al. (1994). Double-masked placebo-controlled trial of precision spectral filters in children who use coloured overlays. *Ophthal. Physl. Opt.* 14, 365–370. doi: 10.1111/j.1475-1313.1994.tb00126.x

Wilkins, A. J., Lewis, E., Smith, F., Rowland, E., and Tweedie, W. (2001). Coloured overlays and their benefit for reading. *J. Res. Read.* 24, 41–64. doi: 10.1111/1467-9817.00132

Wilkins, A. J., Nimmo-Smith, I., Tait, A., McManus, C., Della Sala, S., Tilley, A., et al. (1984). A neurological basis for visual discomfort. *Brain* 107, 989–1017. doi: 10.1093/brain/107.4.989

Wilkins, A. J., Sihra, N., and Myers A. (2005). Increasing reading speed by using colours: issues concerning reliability and specificity, and their theoretical and practical implications. *Perception* 34, 109–120. doi: 10.1068/p5045

Woerz, M., and Maples, W. C. (1997). Test-retest reliability of colored filter testing. *J. Learn. Disabil.* 30, 214–221. doi: 10.1177/002221949703000209

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.

Received: 25 February 2014; accepted: 12 July 2014; published online: 29 July 2014.
*Citation: Uccula A, Enna M and Mulatti C (2014) Colors, colored overlays, and reading skills. Front. Psychol. **5**:833. doi: 10.3389/fpsyg.2014.00833*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 07 May 2014
doi: 10.3389/fpsyg.2014.00398

Is there a bilingual advantage in the ANT task? Evidence from children

Eneko Antón[1]*, Jon A. Duñabeitia[1], Adelina Estévez[2], Juan A. Hernández[3], Alejandro Castillo[4], Luis J. Fuentes[4], Douglas J. Davidson[1] and Manuel Carreiras[1,5,6,7]

[1] Basque Center on Cognition, Brain and Language, Donostia-San Sebastian, Spain
[2] Departamento de Psicología Cognitiva, Social y Organizacional, Faculty of Psychology, University of La Laguna, La Laguna, Spain
[3] Departamento de Psicobiología y Metodología de las Ciencias del Comportamiento, Faculty of Psychology, University of La Laguna, La Laguna, Spain
[4] Departamento de Psicología Básica y Metodología, Faculty of Psychology, University of Murcia, Murcia, Spain
[5] University of Granada, Granada, Spain
[6] Ikerbasque, Basque Foundation for Science, Bilbao, Spain
[7] Departamento de Lengua Vasca y Comunicación, University of the Basque Country EHU/UPV, Bilbao, Spain

Edited by:
Simona Amenta, University of Milano-Bicocca, Italy

Reviewed by:
Petar Milin, Eberhardt Karls University, Germany
Kenneth Robert Paap, San Francisco State University, USA

***Correspondence:**
Eneko Antón, Basque Center on Cognition, Brain and Language, Paseo Mikeletegi 69-2, 20009, Donostia-San Sebastian, Spain
e-mail: e.anton@bcbl.eu

Bilinguals have been shown to outperform monolinguals in a variety of tasks that do not tap into linguistic processes. The origin of this bilingual advantage has been questioned in recent years. While some authors argue that the reason behind this apparent advantage is bilinguals' enhanced executive functioning, inhibitory skills and/or monitoring abilities, other authors suggest that the locus of these differences between bilinguals and monolinguals may lie in uncontrolled factors or incorrectly matched samples. In the current study we tested a group of 180 bilingual children and a group of 180 carefully matched monolinguals in a child-friendly version of the ANT task. Following recent evidence from similar studies with children, our results showed no bilingual advantage at all, given that the performance of the two groups in the task and the indices associated with the individual attention networks were highly similar and statistically indistinguishable.

Keywords: bilingual advantage, inhibitory skills, executive control, attention, ANT task

INTRODUCTION

The so-called "bilingual advantage" (Kroll and Bialystok, 2013), broadly understood as enhanced executive cognitive control for bilinguals as compared to monolinguals, has attracted very much interest in recent years. Different hypotheses have been proposed to account for this bilingual advantage, all of which predict that bilingual individuals will perform better than their monolingual peers in processing incongruent or salient irrelevant information. While there has been considerable evidence to date supporting a bilingual advantage, very recently there has also been an increase in the number of studies showing a similar performance of bilinguals and monolinguals in non-linguistic executive control tasks. The present study provides data collected from a large sample of carefully matched bilinguals and monolinguals suggesting that the so-called bilingual advantage is not generalizable and replicable when the samples are properly controlled.

One of the most commonly studied tasks in which bilinguals have been claimed to outperform monolinguals is the classic Stroop task (Stroop, 1935). In this task, participants have to name the color in which target words are printed. The difference between the latencies to incongruent trials (i.e., the target word to be named is the name of a color and is printed in a different ink color; e.g., the word "green" printed in red color) and the latencies to congruent trials (i.e., target word and its color match; e.g., the word "green" printed in green) is the Stroop effect, and is an index of inhibitory control. The Stroop effect was found to be smaller in bilingual participants than in their monolingual peers and this difference has been claimed to be especially evident in older bilinguals when compared to their monolingual counterparts (e.g., Bialystok et al., 2008; Hernández et al., 2010). However, as we will explain below, recent results have challenged these findings showing negligible differences between bilinguals and monolinguals in the Stroop task (Duñabeitia et al., 2014).

Evidence in favor of the so-called bilingual advantage has been also obtained using the Simon paradigm (Simon and Rudell, 1967). In this task, participants have to respond with either their left or right hand depending on one specific feature of the stimulus (e.g., the color), while ignoring other salient but apparently irrelevant features of the target (e.g., its location). The Simon task includes congruent and incongruent conditions, as a function of the match between the relevant and irrelevant features. The difference between congruent and incongruent trials (the Simon effect) has been typically found to be smaller in bilinguals than in monolinguals (Bialystok et al., 2004). Again, as with the Stroop task, this bilingual advantage has been found to be much stronger in older than in younger adults (Bialystok et al., 2004). However, as in the case of the Stroop task, recent studies have also reported negligible differences between bilinguals and monolinguals in the Simon task (see Prior and MacWhinney, 2010; Humphrey and Valian, 2012; Kousaie and Phillips, 2012; Kirk et al., 2013; Paap and Greenberg, 2013; Sawi and Paap, 2013; Gathercole et al., 2014).

Another task extensively used in the attention domain to show the bilingual advantage is the Attentional Network Test (ANT; Fan et al., 2002). This task, which is a combination of the classic flanker task (Eriksen and Eriksen, 1974) and the cueing task

(Posner, 1980), measures the three independent attentional networks of Orienting, Alerting and Executive control (e.g., Fan et al., 2003). In this task, participants need to respond to the presence of an arrow on the screen, by indicating whether the arrow is pointing to the left or to the right. The critical arrow (e.g., →) can be flanked by another 2 arrows on each side, either pointing in the same direction (Congruent trials; e.g., →→→→→) or in the opposite direction (Incongruent trials; e.g., ←←→←←). Simple lines can also flank the central arrow, this way creating the Neutral condition (e.g., - - → - -). Previous to each flanker trial and after a random time period, participants can be cued about the position where the arrows are going to appear, since the arrows can appear either in the upper or in the lower part of the screen. The Cue factor can be manipulated so that participants see a valid Spatial Cue (i.e., an asterisk in a congruent cueing position), a Double Cue (i.e., one asterisk in the upper part and another one in the lower part), a Neutral Cue (an asterisk in the middle of the screen) or No Cue at all. With the combination of these 4 cue conditions (Double, Spatial, Center and No cue) and 3 flanker conditions (Congruent, Incongruent and Neutral), a measurement of the three attentional networks can be obtained. The index of the Alerting Network can be obtained by subtracting the reaction times in the Double Cue condition and the ones in the No Cue condition. Similarly, the Orienting index can be obtained by comparing the Central Cue and the Spatial Cue conditions. Finally, and possibly the most important for our purposes, the Conflict Effect, which is closely related to executive control, can be obtained by comparing the reaction times to Incongruent and Congruent trials.

In the Revised ANT task (ANT-R, Fan et al., 2009) a fifth cueing condition was created: the Invalid Spatial Cue. This was conceived as the opposite of the Valid Spatial Cue, which precedes the target stimuli in its exact same position. The Invalid Spatial Cue precedes the target arrow in the opposite part of the screen, so that an asterisk in the lower part would precede targets appearing in the upper part of the screen, and an asterisk in the upper part would precede targets appearing in the lower part. By comparing the (longer) latencies to the Invalid Cue condition to the (shorter) reaction times to the Valid Cue trials, the Validity index is obtained, considered as an index of reorienting attention.

The ANT task has been found to show a different developmental pattern for the different networks. Rueda et al. (2004), tested children from 6 to 10 years of age in an adapted version of the ANT task where the arrows were replaced with fishes to make it more child-friendly. Not surprisingly, they found that overall reaction times and error rates decreased gradually as a function of age. When the Alerting, Orienting, and Conflict networks were analyzed separately, the authors found that the developmental pattern was not parallel for these three networks. On the one hand, the Alerting network showed negligible changes between ages 6 and 10. Similarly, the Orienting network failed to show a clear-cut developmental change. In contrast, the Conflict effect showed a remarkable improvement from age 6 to age 7, remaining relatively stable after that.

Similarly to the Stroop and the Simon tasks, when the ANT task has been used to explore differences between bilinguals and monolinguals, an intriguing pattern has been found. For instance, Costa et al. (2008) tested Catalan-Spanish bilinguals and compared them to their monolingual peers. When looking at the specific attention networks, they found that monolinguals showed larger Conflict effects than bilinguals. Besides, in the Alerting network, bilingual participants showed larger benefits than monolinguals due the presence of an Alerting Cue. They also reported that bilingual participants were overall faster than their monolingual peers regardless of the Flanker and Cue type, and they showed that the overall RT differences could not be simply explained by bilinguals just being better than monolinguals at conflict resolution, given that they were also faster in congruent trials. Taken together, these results led them to abandon the hypothesis that the bilingual advantage was the consequence of bilinguals' better ability to process incongruent information, and to propose that it reflected bilinguals' enhanced monitoring abilities.

To further test this hypothesis, Costa et al. (2009) ran a version of the ANT manipulating the monitoring demands using different groups of bilingual and monolingual participants. In a first experiment they created a low-monitoring context, with 92% of the trials belonging to one condition (either Congruent or Incongruent) and 8% to the other condition, thus making the condition of the upcoming target highly predictable. In a second experiment, they created two high-monitoring contexts. In one of the contexts, each condition (i.e., Congruent and Incongruent) was represented by 50% of the trials, making it difficult to predict the condition of the individual trial. In the other context, the authors opted for a 75% congruent-25% incongruent distribution of the trials. Costa et al. found that bilingual participants were overall faster than monolinguals in the highest monitoring context (namely, 50% of the trials per condition), but did not show differences in the magnitude of the Conflict effect. Contrarily, in the low-monitoring context, both groups behaved similarly, with no differences in overall RTs or in the magnitude of the Conflict effect. In the 75–25% context a slight advantage was found in overall RTs and in the Conflict effect for bilinguals, but these effects were modest and exclusively confined to the first experimental block. Hence, the results reported by Costa et al. suggest that (1) the so-called bilingual advantage does not seem to be exclusively related to an enhancement of bilinguals' inhibitory skills (Green, 1998; Bialystok et al., 2004; Kroll et al., 2008; and see also Morales et al., 2013 for an explanation combining inhibitory and monitoring skills), and that (2) the appearance of the bilingual advantage seems to be restricted to certain experimental conditions, often failing in its replication (e.g., Prior and MacWhinney, 2010; Kousaie and Phillips, 2012; Paap and Greenberg, 2013).

Clearly at odds with these findings reported by Costa et al. (2009), a recent study by Pelham and Abrams (2014) testing young adults who were early bilinguals, late bilinguals or monolinguals in the ANT showed a significant bilingual advantage in conflict resolution. They found that monolinguals were slower than the two bilingual groups in incongruent trials, showing larger conflict effects than both late and early bilinguals (with no differences between the last two).

Although the main focus of bilingualism research using the ANT task has been the Conflict effect, given its direct relationship

with executive control and its implications for the bilingual advantage based on inhibitory skills; it is worth noting that there has also been evidence of differences in the Alerting effect (Costa et al., 2008; but see Costa et al., 2009) and in the Orienting network (Colzato et al., 2008; but see Hernández et al., 2010). Clearly, it is difficult to extract a take-home-message from the bulk of evidence gathered from ANT studies with bilingual and monolingual adult samples, given the high degree of variance in the observed results.

Leaving aside the debate about critical experimental settings, tasks or contexts that lead to the appearance or vanishing of the bilingual advantage, it is worth noting that the strongest pieces of evidence supporting it come from adult research and especially from research done with elder adults. However, this bilingual advantage is more elusive in research with children and the number of discrepant studies of this type has increased in recent years. Curiously, it should be mentioned that even researchers showing differences between bilingual and monolingual adults admit that the evidence in favor of a bilingual advantage in children is certainly limited (Bialystok et al., 2010, 2012; see also Hilchey and Klein, 2011, for a review). Furthermore, it has been suggested that some factors other than the mere linguistic profile of the participants may play a very important role in the emergence of the bilingual advantage in different tasks. For instance, Morton and Harper (2007) tested a group of bilingual and monolingual children in a Simon task and they found no differences in their performance as a function of the number of languages they knew. Instead, they found a significant correlation between their socio-economic status (SES) and their performance in the task, arguing that the SES, not bilingualism, was the crucial factor in producing the effect. Hence, the number of intra-experimental and external factors that seem to have a direct impact on the appearance (and the magnitude) of the bilingual advantage is increasing, and the true nature of bilingual outperformance in executive control tasks remains unclear, casting doubts on some of the claims that have lead the field in the last decade. In this line, Paap and Greenberg (2013) recently reported that the studies which have failed to obtain a bilingual advantage should not be ignored. They noted that many of the studies showing a bilingual advantage could possibly be showing a Type I error, due to inadequately matched or very small groups, uncontrolled external factors or task-dependency effects. They concluded that the replicability and the cross-study reliability of this advantage are markedly low.

Following this line of reasoning, in a recent study, Duñabeitia et al. (2014) compared the performance of a group of more than 250 bilingual children to that of a group of very well matched monolinguals in both the classic Stroop task and the Numerical Stroop task (a variation of the classic task with minimal involvement of language). Following the claims raised, among others, by Paap and Greenberg (2013) and Morton and Harper (2007), Duñabeitia et al. carefully matched participants for age, reading and mathematical abilities, and verbal and non-verbal IQ, together with some socio-economic indicators. In a series of different analyses, Duñabeitia et al. found no signs of a difference in the performance of these two groups. These findings lead the authors to conclude that the so-called bilingual advantage in executive control tasks seems to be inexistent in children. Nonetheless, as they acknowledged, further research is needed in order to shed light on the replicability of the bilingual advantage across tasks.

These conclusions are also endorsed by a recent study by Gathercole et al. (2014), who tested a large number of Welsh children and adults in different tasks ($n = 650$ in a card sorting task, $n = 557$ in the Simon task and $n = 354$ in a grammaticality judgment task). The different groups tested included English monolinguals and bilinguals coming with different degrees of use of Welsh and English (i.e., bilinguals who only spoke Welsh at home, bilinguals who used both Welsh and English at home, and bilinguals coming from English-speaking homes). Importantly, Gathercole et al. found no evidence for a bilingual advantage. No differences were found in the switch cost or overall performance in the card sorting task. Similarly, negligible differences were found in the Simon task. The grammaticality judgment task also failed to reveal any systematic bilingual advantage.

Considering the lively debate about how bilingualism may affect performance in the ANT task, in the current study we tested a group of 360 children (180 bilinguals, 180 monolinguals) of different ages in a child-friendly version of the ANT (see Rueda et al., 2004). Similarly to the careful matching of the participants tested in the study by Duñabeitia et al. (2014), special care was taken to avoid the influence of uncontrolled factors in the data observed. Following the inconsistent results obtained in the ANT with adult participants (see Costa et al., 2009; Pelham and Abrams, 2014), the absence of a bilingual advantage in the study with children presented by Duñabeitia et al. (2014), and the results reported by Rueda et al. regarding the different development of the attention networks as a function of age, here we investigated (1) whether there is a bilingual advantage in children in any of the attention networks, and (2) whether the development of these networks is similar or different for bilingual and monolingual children.

METHODS

PARTICIPANTS

Two groups of participants were recruited from different schools in Spain ($n = 360$, females $= 211$). The first group was made up of 180 Spanish monolingual children (females $= 106$) from second, third, fourth and fifth grades of elementary school and grade one from secondary school. These monolinguals were recruited from Spanish schools in places where Spanish is the only official language, and none of them had fluent knowledge of any other language than Spanish. Also, none of them corresponded to any immigrant minority and they were only exposed to Spanish at home. The second group was formed by 180 bilingual children (females $= 105$) from the same grades who were born and lived in the Basque Country. The Basque Country is a Spanish region where two languages, Basque and Spanish, are co-official. All these bilingual children were attending bilingual schools where both languages were used as vehicular languages. According to the legal requirements, bilingual schools in the Basque Country ensure that teachers switch from one to the other language as they switch academic subjects, making sure of a similar distribution of the languages across subjects and school time (50% in each language). This way, Basque children attending bilingual schools

are exposed actively to the two languages on a daily basis during schooling. A linguistic competence questionnaire filled in by 171 of the 180 bilingual children's parents (namely, 95% of the sample) showed that bilingual participants had acquired the two languages very early in life, with overall age-of-acquisition scores of 0.58 years ($SD = 0.77$) for Spanish and of 2.23 years ($SD = 1.07$) for Basque. The parents' subjective ratings for the children's performance in Basque and Spanish were collected on a 0-to-10 scale, where 10 represented a perfect knowledge and use of the language. Children's mean proficiency scores in Spanish was 8.65 ($SD = 1.17$), and their score in Basque was 5.96 ($SD = 1.63$).

The reason for selecting samples of children instead of adult samples is twofold. First, considering the idiosyncrasy of the bilingual educational system in the Basque Country (see above), a relatively high degree of control of children's use of the two languages can be applied. Simply by checking their academic syllabus and the language in which each subject is being taught, daily exposure to both languages can be ensured. And second, considering that the most reliable pieces of evidence supporting the so-called bilingual advantage have been obtained for individuals that are not at ceiling level in their executive functions (e.g., elderly), it could be tentatively suggested that any difference between bilinguals and monolinguals should also emerge in samples of individuals who have not reached yet a fully developed attentional system (e.g., children). The different cognitive and executive skills develop progressively during childhood, and while some of them are relatively mature around age 12–13, many other executive processes are only fully developed or established during mid-adolescence or adulthood (see Anderson, 2002, for review).

In order to explore the developmental trajectory of the attention networks, we divided the sample of bilinguals and monolinguals into three evenly distributed subgroups. Monolingual and bilingual 2nd and 3rd graders were classified as Group 1, 4th and 5th graders were classified as Group 2, and 6th graders and students from the first high school grade were classified as Group 3. 120 children were included in each group, half of them ($n = 60$) corresponding to a monolingual environment and the other half corresponding to a bilingual context. Pairwise comparisons within each group showed no differences (all $ps > 0.11$) between bilinguals and monolinguals in age, gender, overall reading and arithmetic skills (as assessed by their teachers on a 1-to-5 Likert scale), verbal, non-verbal and composed IQ [obtained from the Spanish version of the Kaufman Brief Intelligence Test (1990), K-BIT], income at home (classified according to the following categories: >3000€/month, category 1; 2001–3000€, category 2; 1601–2000€, category 3; 1201–1600€, category 4; 750–1200€, category 5 and <750€ category 6), number of years of formal education of the parents, and parental work status (including three possible categories: neither works, only one of them works, both of them work). Furthermore, we made sure that none of the participants had any specific developmental, psychological, psychiatric, or educational disorder, deficit, or special need by including a series of questions in this regard in the questionnaires completed by parents and teachers. Besides, none of the children had repeated any academic year and no child with scores below the 20th centile in verbal, non-verbal, and combined IQ tests was included in the sample. Hence, the two groups were carefully matched in many socio-economic and cognitive measures (see **Table 1** for detailed comparisons).

DESIGN

In this version of the child Attention Network Test (ANT) two within-subject factors were manipulated, Cue type (Double Cue, Valid Cue, Invalid Cue, Neutral Cue and No Cue) and Flanker type (Incongruent, Congruent), leading to a total of 10 conditions. As already explained in the Introduction, Fan et al. (2009) suggested that the inclusion of an index of validity within the cueing conditions provides an additional measure of the ability to reorient attention. Hence, valid and invalid cues were included in the current design too. The cueing manipulations were created by presenting (or not) an asterisk on the screen prior to the presentation of the target strings. These cues could be presented at the same position of the upcoming target (Valid condition), or in the opposite position (Invalid condition). In order to create the Double Cue condition, two asterisks were presented at the same time above and below the center of the screen. The Neutral Cueing condition was created by presenting the asterisk at the center of the screen, and the No Cue condition was created by not providing any visual cue. Regarding the flanker manipulation, the target was a left- or right-pointing yellow fish (1.6°), presented above or below the fixation cross. This central fish was flanked on both sides by two fishes pointing either in the same direction (Congruent trials), or in the opposite direction (Incongruent trials). The distance between the fishes was 0.21°. The target and flankers subtended 8.84° and were presented 1° above and below the fixation cross over a blue-green background. For detailed description of the stimuli and procedure, see Rueda et al. (2004).

PROCEDURE

All the stimuli were presented on a computer screen. Each trial began with a fixation cross (1° of visual angle) with a random duration between 400 and 1600 ms. Then a cue (an asterisk) could appear in any of its variants (see below) for 150 ms. Next, a centered fixation cross appeared on the screen for 450 ms, immediately followed by the target and flanker stimuli. The target string stayed on the screen until a response was given or for a maximum of 1700 ms. After each trial, feedback was provided.

A session of the ANT consisted in a total of 288 trials. Each trial represented one of the 10 conditions mentioned above (Cue type × Flanker type). To keep the high-monitoring demanding context, 50% of the trials belonged to the Congruent condition and the other 50% to the Incongruent condition. Regarding each cueing condition, there were 72 Double Cue, 48 Valid, 48 Invalid, 48 Neutral Cue and 72 No Cue trials. Participants were seated at a distance of about 55 cm from the screen and they were instructed with a series of practice trials to indicate the direction of the central fishes of the strings, pressing the "L" key in the keyboard for right responses or the "S" key for left responses. Both accuracy and reaction times were recorded in each experimental trial.

DATA ANALYSIS

Reaction times below 200 ms (only representing 0.12% of the data) were excluded. Reaction time data was trimmed by using

Table 1 | Characteristics of the samples tested in the experiment.

Age group	Language group	Age (in years)		Reading scores (1–5)		Math scores (1–5)		Verbal IQ (centiles)		Non-verbal IQ (centiles)		General IQ (centiles)		Incomes (category)		Parents' education (years)		Parents' work situation (category)	
		Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Group 1 Primary 2nd and 3rd	Bilinguals	7.57	0.59	4.53	1.17	4.52	0.93	77.18	14.58	63.00	22.31	68.82	17.88	1.98	1.07	14.30	2.49	1.90	0.35
	Monolinguals	7.55	0.53	4.57	0.98	4.57	0.87	79.28	15.76	60.85	22.18	69.73	19.74	2.15	0.99	13.88	2.76	1.90	0.35
	p value	0.88		0.84		0.72		0.31		0.48		0.70		0.25		0.29		1.00	
Group 2 Primary 4th and 5th	Bilinguals	9.53	0.57	4.75	0.95	4.87	0.89	63.72	18.62	66.13	18.43	62.30	17.56	1.77	0.96	14.59	2.16	2.00	0.00
	Monolinguals	9.50	0.60	4.78	0.83	4.82	0.87	65.32	19.12	66.53	17.81	63.32	17.13	1.88	0.94	14.44	2.39	2.00	0.00
	p value	0.78		0.84		0.75		0.65		0.90		0.76		0.55		0.71		1.00	
Group 3 Primary 6th and Secondary 1st	Bilinguals	11.43	0.65	4.57	1.06	4.42	0.91	56.93	18.23	68.03	17.90	59.52	17.64	1.48	0.68	14.62	2.30	1.92	0.28
	Monolinguals	11.47	0.54	4.58	0.91	4.63	0.84	61.20	17.73	63.10	19.78	59.37	19.28	1.65	0.66	14.07	2.34	1.95	0.22
	p value	0.73		0.92		0.13		0.12		0.11		0.96		0.17		0.18		0.42	
Total	Bilinguals	9.51	1.69	4.62	1.06	4.60	0.93	65.94	19.11	65.72	19.64	63.54	18.02	1.74	0.93	14.50	2.31	1.94	0.26
	Monolinguals	9.51	1.70	4.64	0.91	4.67	0.86	68.60	19.14	63.49	20.03	64.14	19.14	1.89	0.89	14.13	2.50	1.95	0.24
	p value	0.93		0.79		0.42		0.16		0.31		0.77		0.13		0.12		0.66	

the classic 2.5SD criterion, resulting in the exclusion of the 2.49% of the data, and RTs associated with erroneous responses were not included in the latency analyses. Before focusing on the individual indices for each attention network, all the conditions were analyzed in a general ANOVA including Cue Type (No Cue, Valid Cue, Invalid Cue, Double Cue and Neutral Cue) and Flanker Type (Congruent and Incongruent) as within-participant factors, and Language (Bilinguals and Monolinguals) and Group (First, Second and Third group) as between-participants factors. In subsequent analyses we looked at the different attention networks by measuring the following indexes: the difference between Congruent and Incongruent trials as a reflection of executive control (Conflict effect), the differences between the Double Cue and the No Cue conditions for the alerting network (Alerting effect), the orienting network as measured by the difference between the trials with a Neutral Cue and trials with a Valid Cue (Orienting effect), and finally the difference between the trials with a Valid Cue vs. the trials with an Invalid Cue as markers of the Validity effect. Detailed information about the RT and error data is presented in **Table 2**.

RESULTS

GENERAL ANALYSES

In the RT analysis, we found significant main effects of Flanker Type [$F_{(1,\ 354)} = 1624.68$, $MSE = 1993.35$, $p < 0.01$], Cue Type [$F_{(4,\ 1416)} = 237.19$, $MSE = 1298.75$, $p < 0.01$] and Group [$F_{(2,\ 354)} = 120.07$, $MSE = 66486.08$, $p < 0.01$]. In contrast, the main effect of Language was not significant [$F_{(1,\ 354)} = 2.22$, $MSE = 66486.08$, $p > 0.13$]. The 2-way interaction between Flanker Type and Group was significant [$F_{(2,\ 354)} = 12.5$, $MSE = 1993.35$, $p < 0.01$], and the same was true for the interaction between Flanker Type and Cue Type [$F_{(4,\ 1416)} = 24.12$, $MSE = 893.76$, $p < 0.01$]. None of the other interactions was significant.

In error rate analysis, both Language groups performed similarly ($F < 1$). The main effects of Flanker Type [$F_{(1,\ 354)} = 303.20$, $MSE = 35.25$, $p < 0.01$], Cue Type [$F_{(4,\ 1416)} = 11.52$, $MSE = 17.61$, $p < 0.01$], and Group [$F_{(2,\ 354)} = 43.53$, $MSE = 210.73$, $p < 0.01$] were significant. The only significant interactions found were the Flanker Type * Group interaction [$F_{(2,\ 354)} = 6.85$, $MSE = 35.25$, $p < 0.01$], and the Flanker Type * Cue Type interaction [$F_{(4,\ 1416)} = 90.32$, $MSE = 17.44$, $p < 0.01$].

Thus it is important to notice that none of the interactions with Language were significant, showing that the same effects hold for bilinguals and monolinguals.

THE THREE ATTENTIONAL NETWORKS

Considering the reliable Flanker Type * Cue Type interactions, and following preceding research, we explored each of the effects mentioned above individually (i.e., Conflict, Alerting, Orienting and Validity), and the manner in which the between-participants factors Group and Language could modulate them (see **Table 3** and **Figure 1** for comparisons between Language groups; and see **Table 4** and **Figure 2** for a detailed comparison between Language Groups in each Age Group).

Executive network: the Conflict effect

In the RT analysis, the Conflict effect as measured by the factor Condition (Congruent vs. Incongruent trials) was significant [$F_{(1,\ 354)} = 1624.68$, $MSE = 398.67$, $p < 0.01$], as well as the main effect of Group [$F_{(2,\ 354)} = 120.07$, $MSE = 13297.22$,

Table 2 | Reaction times and error rates to each condition.

	Conditions															
	Double Cue		Neutral Cue		Valid Cue		Invalid Cue		No Cue		Congruent		Incongruent		Total	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
REACTION TIMES																
Bilinguals	690.30	110.13	705.87	112.55	672.67	107.13	724.39	103.20	714.49	108.34	670.88	104.01	732.21	109.47	701.55	105.75
Monolinguals	676.12	101.31	692.86	111.44	659.59	106.71	711.09	108.94	703.99	105.34	659.41	103.97	718.05	106.95	688.73	104.48
ERROR RATES																
Bilinguals	4.92	5.40	4.92	5.41	4.56	5.30	5.91	6.30	5.69	5.61	3.45	4.58	6.95	5.90	5.20	4.91
Monolinguals	4.58	5.64	4.99	5.72	4.20	5.84	5.60	6.43	5.02	5.68	3.18	4.78	6.57	6.30	4.88	5.28

Table 3 | Attentional networks, measured as the difference in reaction times and error rates.

	Attentional networks							
	Conflict index		Orienting index		Alerting index		Validity index	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
REACTION TIMES								
Bilinguals	61.34	29.51	33.20	39.85	24.19	32.82	51.72	41.95
Monolinguals	58.64	28.94	33.27	38.50	27.87	30.76	51.50	42.78
ERROR RATES								
Bilinguals	3.50	3.91	0.36	4.22	0.76	4.00	1.35	4.51
Monolinguals	3.39	3.71	0.79	3.77	0.43	4.09	1.40	4.28

$p < 0.01$] and the interaction between them [$F_{(2, 354)} = 12.50$, $MSE = 398.67$, $p < 0.01$]. It took longer for participants to respond to the Incongruent trials as compared to the Congruent ones, and participant speed of response increased as a function of age (see below). Importantly, the main effect of Language was not significant [$F_{(1, 354)} = 2.22$, $MSE = 13297.22$, $p > 0.13$], and it did not interact with Condition ($F < 1$) or with Group ($F < 1$). The 3-way Language*Condition*Group interaction was not significant [$F_{(2, 354)} = 2.22$, $MSE = 398.67$, $p > 0.11$]. Hence, we can conclude that monolinguals and bilinguals showed highly similar Conflict effects.

In order to explore the origin of the significant Condition*Group interaction, follow-up contrasts were run collapsing the data across linguistic profiles. Pairwise contrasts showed that the differences in the responses to the two types of flankers (Congruent, Incongruent) decreased with age. Thus, when comparing the Conflict effect in each Group, we observed that the first group showed the largest Conflict effect (average of 70 ms), and that this effect progressively diminished with age (Group 2 = 57 ms; Group 3 = 52 ms). Pairwise tests showed that the effect was significantly larger for Group 1 than for Group 2 and Group 3 [Group 1 vs. Group 2: $t_{(238)} = 3.18$, $p < 0.01$; Group 1 vs. Group 3: $t_{(238)} = 4.54$, $p < 0.01$], while the difference was not significant between Groups 2 and 3 [$t_{(238)} = 1.70$, $p < 0.1$].

In error rate analysis only the main effects of Condition [$F_{(1, 354)} = 303.20$, $MSE = 7.05$, $p < 0.01$] and Group [$F_{(2, 354)} = 43.53$, $MSE = 42.15$, $p < 0.01$] were significant.

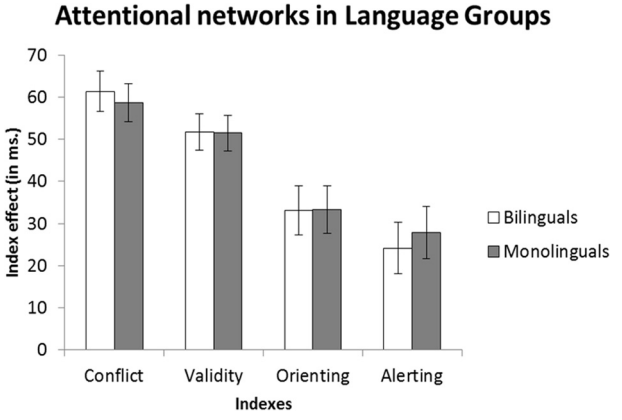


FIGURE 1 | Comparison of indexes across Language Groups.

The only significant interaction was found between Condition and Group [$F_{(1, 354)} = 6.85$, $MSE = 7.05$, $p < 0.01$]. Replicating the RT data, the error data showed a clear Conflict effect, with higher error rates in incongruent than in congruent conditions and a modulation of the percentages of errors as a function of age (i.e., overall error rates diminished as a function of age). Given the significant interaction, we can conclude that the magnitude of the Conflict effect decreased as a function of age. Importantly, the Language effect and the interactions between this and the other factors were negligible (all Fs < 1 and all ps > 0.5).

Alerting network: the Alerting effect

When considering the differences in RTs between the Double Cue and the No Cue conditions, only the main effects of Condition [$F_{(1, 354)} = 239.44$, $MSE = 509.37$, $p < 0.01$] and Group [$F_{(2, 354)} = 118.55$, $MSE = 13364.56$, $p < 0.01$] were significant. The Language effect was not significant [$F_{(1, 354)} = 2.05$, $MSE = 13364.56$, $p > 0.15$]. None of the interactions were significant (Fs < 1.20, ps > 0.27]. Hence, participants responded faster to Double Cue trials than to No Cue trials and they became overall faster as their age increased but the difference between the cueing conditions did not differ across ages or across language profiles.

In the error rate analysis, the only significant effects corresponded to the factors Condition [$F_{(1, 354)} = 7.81$, MSE=8.25, $p < 0.01$] and Group [$F_{(2, 354)} = 41.25$, $MSE = 44.43$, $p < 0.01$], showing that participants made more errors in No Cue trials than in Double Cue trials and that the number of errors decreased as a function of age. No other effects or interactions were significant (all Fs < 1.1 and all ps > 0.3).

Orienting network: the Orienting effect

The Orienting effect (i.e., Valid Cue vs. Neutral Cue) was significant [$F_{(1, 354)} = 260.30$, $MSE = 763.89$, $p < 0.01$], as was the main effect of Group [$F_{(2, 354)} = 109.45$, $MSE = 14488.40$, $p < 0.01$]. Responses to trials with a Valid Cue were faster than responses to trials with a Neutral Cue and averages RTs decreased as a function of age. In contrast, the main effect of Language was not significant [$F_{(1, 354)} = 2.12$, $MSE = 14488.40$, $p > 0.14$], and none of the interactions involving the factor Language was significant (all Fs. < 1). A marginal interaction between Condition and Group was found [$F_{(2, 354)} = 2.84$, $MSE = 763.89$, $p < 0.07$], suggesting that the magnitude of the Orienting effect decreased with age. Follow-up pairwise contrasts showed similar Orienting effects for Groups 1 and 2 (39 and 34 ms, respectively; $t < 1$), and a significantly smaller effect for Group 3 [27 ms; Group 1 vs.

Table 4 | Latency differences in attentional networks in each age group.

	Conflict effect				Orienting effect				Alerting effect				Validity effect			
Age Group	**Bilinguals**		**Monolinguals**		**Bilinguals**		**Monolinguals**		**Bilinguals**		**Monolinguals**		**Bilinguals**		**Monolinguals**	
Group 1	73.53	(36.21)	66.63	(35.81)	38.02	(49.51)	39.59	(49.61)	21.71	(41.38)	30.06	(41.01)	51.56	(50.14)	56.76	(47.72)
Group 2	54.44	(21.51)	60.60	(26.15)	34.12	(38.12)	33.95	(35.92)	23.77	(32.63)	25.18	(20.58)	54.47	(43.28)	58.27	(43.66)
Group 3	56.04	(25.29)	48.69	(20.12)	27.47	(29.25)	26.26	(25.62)	27.09	(21.82)	28.37	(27.44)	49.14	(30.63)	39.49	(33.88)

Means and SD (in parenthesis) are displayed.

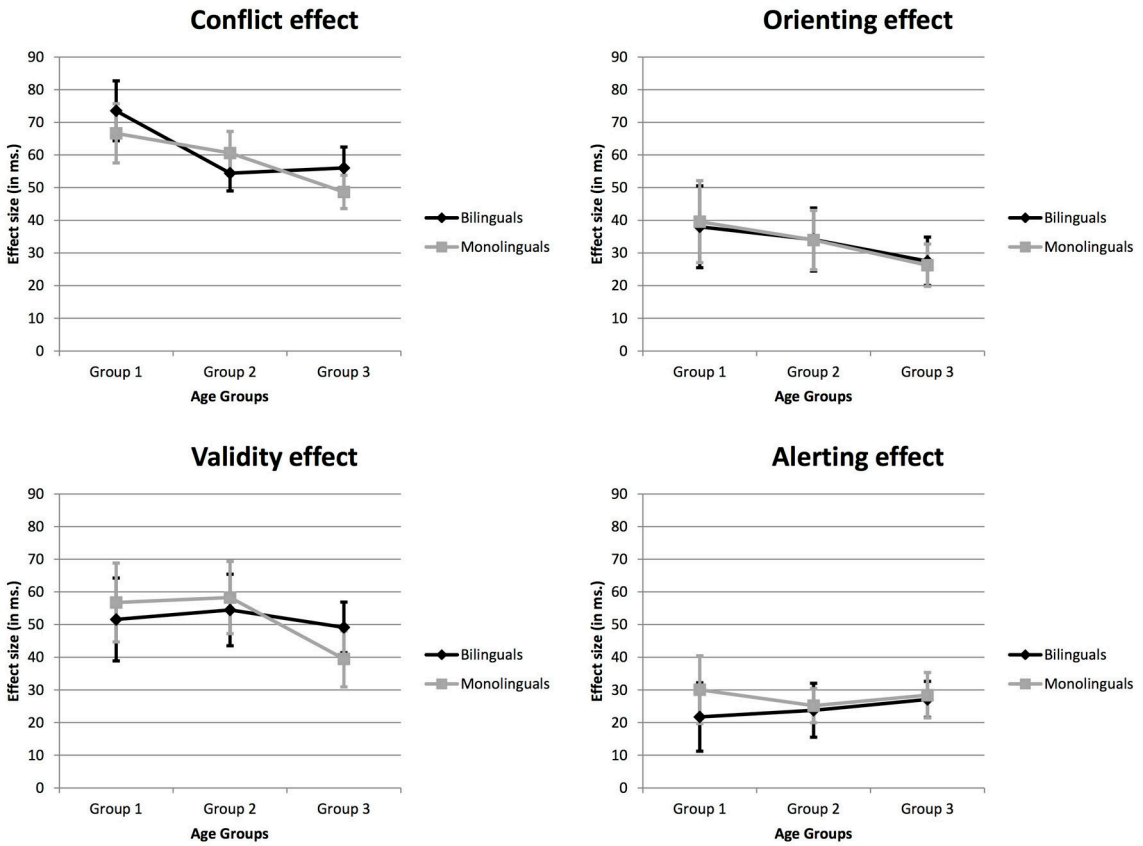


FIGURE 2 | The four indexes, representing the attentional networks, across age groups and language groups.

Group 3: $t_{(238)} = 2.32$, $p < 0.03$; Group 2 vs. Group 3: $t_{(238)} = 1.71$, $p < 0.09$].

In the error rate analysis, the only significant effects found were in Condition [$F_{(1,\ 354)} = 7.33$, $MSE = 8.06$, $p < 0.01$], showing more errors in the Neutral Cue condition than in the Valid Cue condition, and Group [$F_{(2,\ 354)} = 34.74$, $MSE = 45.66$, $p < 0.01$], showing a decrease in the amount of errors as a function of age. No other effects or interactions were significant (all Fs < 1.1 and all ps > 0.3).

Reorienting: the Validity effect

The difference between trials with a Valid Cue and trials with an Invalid Cue were significant in the RT analysis [main Condition effect: $F_{(1,\ 354)} = 539.92$, $MSE = 888.06$, $p < 0.01$], and the Group effect was also significant [$F_{(2,\ 354)} = 117.92$, $MSE = 13211.03$, $p < 0.01$]. Invalid Cues produced longer response times than Valid Cues, and the overall response times decreased as a function of age. These two effects marginally interacted with each other [$F_{(2,\ 354)} = 2.78$, $MSE = 888.06$, $p < 0.07$], suggesting that the magnitude of the Validity effect decreased with age. Follow-up t-tests showed that the magnitude of the Validity effect was similar for Groups 1 and 2 (54 and 56 ms, respectively; $t < 1$), and that the effect was smaller for Group 3 (44 ms) than for Group 2 [$t_{(238)} = 2.44$, $p < 0.02$] and, although marginally significant, than for Group 1 [$t_{(238)} = 1.84$, $p < 0.07$]. Critically, the main effect of Language was not significant [$F_{(1,\ 354)} = 2.37$, $MSE = 13211.03$, $p > 0.12$], and none of the interactions involving the Language factor were significant either (all Fs < 1.15 and ps > 0.32).

Parallel findings were also observed in the error rate analysis, showing significant Condition [$F_{(1,\ 354)} = 35.60$, $MSE = 9.59$, $p < 0.01$] and Group effects [$F_{(2,\ 354)} = 37.15$, $MSE = 51.80$, $p < 0.01$], together with a marginal interaction between these two factors [$F_{(2,\ 354)} = 3.03$, $MSE = 9.59$, $p < 0.06$]. Again, no other effects or interactions were significant (all Fs < 1 and all ps > 0.5).

BAYESIAN NULL HYPOTHESIS TESTING

Given that classical hypothesis testing does not allow for accepting the null hypothesis, we tested the critical differences of interest following a Bayesian approach (see Rouder et al., 2009, among others). For each index (Conflict, Validity, Orienting and Alerting), we used a Bayes factor (BF) approach to compare a model that assumed no differences between bilinguals and monolinguals (H0) against a model that assumed that bilinguals perform differently from monolinguals (H1). With this test, the null hypothesis is accepted if the resulting BF is below 0.3, and the alternative hypothesis is accepted if it is above 3 (see Kruschke, 2011, **Figure 3** in page 6). When comparing bilinguals and monolinguals' Conflict effects, results favored the acceptance of the null model ($BF < 0.18$). The other three attentional

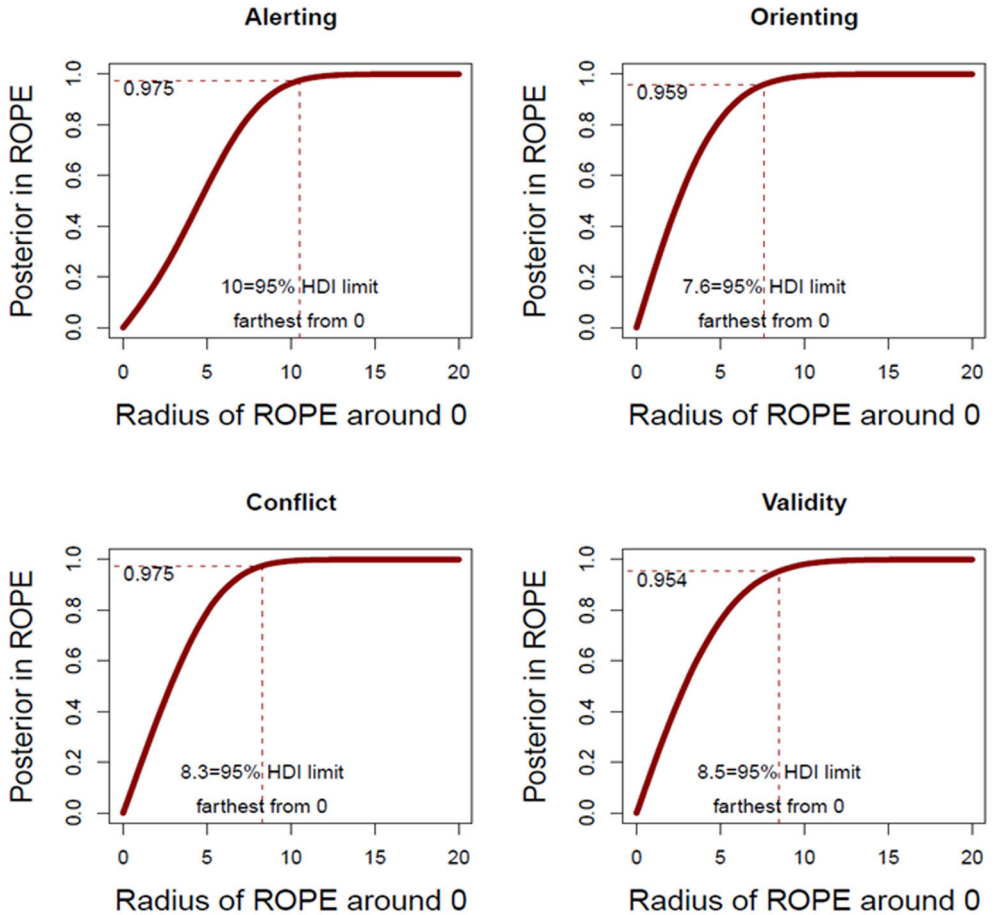


FIGURE 3 | Proportion of the posterior distribution falling within the ROPE as a function of ROPE width. X axis shows how far the ROPE limit is from 0 value (no differences). Y axis reflects the proportion of the posterior distribution that falls inside the ROPE. Dotted line shows the proportion at the right edge of the highest posterior density interval (HDI).

networks responded similarly, all of them being better explained by a null model as compared to the alternative ($BF < 0.12$ for the Orienting effect, $BF < 0.21$ for the Alerting effect, and $BF < 0.13$ for the Validity effect). These results suggest support for the hypothesis of no difference.

We further explored the reliability of the current lack of differences using Bayesian Parameter Estimation by testing the degree of confidence of the null value with the Region of Practical Equivalence (i.e., ROPE; see Kruschke, 2013, for details). Following this approach, a ROPE comprising the range of values assumed to be statistically equal to the null value (i.e., how much of a difference is accepted to be considered equal to no differences at all) is determined by previous findings in the field. If at least 95% of the posterior distribution (i.e., the prior distribution updated by the distribution of the current data) falls within this ROPE, the null hypothesis should be accepted. In contrast, if 95% of the posterior distribution falls outside the ROPE, then the alternative hypothesis should be accepted. The ROPE width would ideally be taken from preceding similar studies, but in this case there is not a consensus in the literature about the smallest meaningful difference. Therefore, we calculated the proportion of the posterior falling within the ROPE boundaries for a range of ROPE limits from 0 to 20 ms in each index (Conflict, Alerting, Orienting and Validity). This approach allows us to calculate the range of values surrounding 0 that should be accepted as equivalent to no-differences (i.e., the ROPE) to accept the null hypothesis. As seen in **Figure 3**, in order to get the 95% or more of our posterior distributions within the ROPEs, the radii of the ROPEs need to be set to values ranging from 7.6 to 10 ms (10 ms for Alerting, 7.6 for Orienting, 8.3 for Conflict and 8.5 for Validity). In essence, this means that, if we accept differences between 7.6 and 10 ms as equivalent to no differences at all, and given that then the majority of the distribution of the differences falls below these limits, we take this as support for the null hypothesis([1]). Considering that the differences found between bilinguals and monolinguals in the four indices are far below these cutoff points (4 ms for Alerting,

[1]It should be noted that this does not imply that any between-group difference larger than 10 ms would significantly allow us to accept the alternative hypothesis. For this to be the case, at least 95% of the posterior distribution should lay outside the ROPE, and this would necessarily imply a much larger difference.

0 ms for Orienting, 3 ms for Conflict and 0 ms for Validity), considering also that reliable differences in RTs of 10 ms in studies of children are rarely reported (note also that even in adult differences of 10 ms in the conflict effect between bilinguals and monolinguals may result in a non-significant effect (see Costa et al., 2009), we believe the data support the null hypothesis (no differences between bilinguals and monolinguals).

GENERAL DISCUSSION

The aim of this study was to investigate whether bilingual children exhibit an advantage as compared to their monolingual peers in the ANT task, which has been typically considered the paradigm best suited to explore the different attention networks. As described in the Introduction, different explanations have been given for the so-called bilingual advantage (see Green and Abulatebi, 2013; Kroll and Bialystok, 2013); but all of them coincide in suggesting that the continuous use and control of (and switching between) two languages provides bilinguals with a set of enhanced attention skills that ultimately leads to the emergence of differences between monolinguals and bilinguals in different non-linguistic tasks closely associated with executive control. In light of some recent studies failing to replicate the bilingual advantage with different populations (e.g., Paap and Greenberg, 2013; Duñabeitia et al., 2014; Gathercole et al., 2014), and considering the existing debate between researchers suggesting that bilinguals outperform monolinguals in the ANT task (e.g., Kapa and Colombo, 2013; Pelham and Abrams, 2014) and those suggesting that the bilingual advantage in this task is restricted to certain conditions and designs (e.g., Costa et al., 2009), we investigated whether a large sample of bilingual children would exhibit better performance in this task than a group of carefully matched monolingual children. Our results unambiguously demonstrated that the so-called bilingual advantage could not be replicated in the ANT when a sufficiently large and well-matched group of bilingual and monolingual children were tested.

Our results add to a growing body of evidence showing that most forms of bilingual advantage in tasks exploring attention skills may well be the result of uncontrolled factors (e.g., Morton and Harper, 2007; Paap and Greenberg, 2013; see also Paap and Liu, 2014, and Paap, submitted, for review) or specific conditions associated with the design and procedure (e.g., Costa et al., 2009). Also, together with the results provided by Duñabeitia et al. (2014) from a large-scale study testing monolingual and bilingual children in two different versions of the Stroop task and by Gathercole et al. (2014), who tested a large number of Welsh-English bilinguals and English monolinguals in different tasks, these results demonstrate the clear similarity between monolingual and bilingual children in their performance in tasks with high executive control demands.

We argue that if the so-called bilingual advantage were a consequence of bilinguals' enhanced inhibitory skills, a reduced Conflict effect should have been found for the bilingual group (i.e., smaller differences between Incongruent and Congruent trials for bilinguals than for monolinguals). This was not the case, and participants performed in a highly similar fashion in these two conditions regardless of their linguistic profile. On the other hand, if the previously reported bilingual advantage were the result of bilinguals' enhanced monitoring skills, one would have expected an overall difference between groups in the RTs and/or in the error rates (e.g., Costa et al., 2009; see also Wu and Thierry, 2013), but again we did not find any supporting data for this claim (see also Duñabeitia et al., 2014, for similar results).

It is worth mentioning that the lack of a bilingual advantage in this study cannot be ascribed to a general lack of sensitivity of our design to the specific attention network(s) that may underlie such a difference between bilinguals and monolinguals. Replicating preceding evidence from the monolingual domain, we have shown that bilingual and monolingual children exhibited longer latencies and higher error rates for Incongruent trials than for Congruent trials (namely, a significant Conflict effect). Similarly, a better performance of both groups was found in the Double Cue trials as compared to the No Cue trials (namely, a significant Alerting effect). Also, participants' responses to the Valid Cue trials were faster and more accurate than their responses to Central Cue (i.e., a significant Orienting effect). Finally, participants showed longer RTs and higher error rates in trials involving an Invalid Cue than in trials with a Valid Cue (i.e., a significant Validity effect). Hence, considering that the current results fully replicate the indices observed in preceding studies with the ANT task (e.g., Fan and Posner, 2004; Fan et al., 2005; Wang and Fan, 2007; Ishigami and Klein, 2010; Yin et al., 2012; Mackie et al., 2013 among many others), it is hardly possible that potential differences between bilinguals and monolinguals were masked due to a lack of statistical power of the current study (see also the magnitude of the *F*-values at this regard). Furthermore, from a developmental point of view, the current study has replicated and extended the findings observed by Rueda et al. (2004) in a similar study testing a smaller group of monolingual children. The same developmental trend observed in that study can be seen here, suggesting that the Conflict effect (hence, the executive network) is the attentional index that is most sensitive to a developmental change, greatly diminishing as a function of age. On the other hand, we see more modest changes in the Validity and Orienting effects (note that the interactions were marginally significant in spite of the sample size), and no significant changes in the Alerting effect as a consequence of age.

In a nutshell, and in spite of the statistical power of the current study, no significant differences between bilingual and monolingual children emerged in their performance in the ANT task. Furthermore, when taking the Bayesian approach to test the null hypothesis against the alternative, the null appears as the strongest candidate. When the analysis was based on the ROPE approach, we also found support for the null hypothesis. In this analysis we found limits for the difference between groups that were in fact larger than previously reported differences in adults.

Certainly, we want to avoid generalizing the observed lack of bilingual advantage to other age groups, and as already discussed in Duñabeitia et al. (2014), our claims are exclusively endorsing the conclusion that the so-called bilingual advantage in tasks focusing on participants' attention skills is inexistent, or at best, extremely inconsistent and elusive. As discussed in the Introduction, both behavioral and neuroimaging evidence (see, among many others, Luk et al., 2010; Gold et al., 2013) suggest some form of bilingual advantage in similar tasks with adult

samples. Hence, as mentioned by Kroll and Bialystok (2013), the existence of a bilingual advantage in adulthood cannot be ignored, even though the degree to which those findings can be generalized to all adult bilingual samples is limited (see Paap and Greenberg, 2013, among others). It should be considered that the so-called bilingual advantage may emerge as a consequence of lifelong bilingualism mainly in later stages of life (e.g., the elderly).

Leaving aside the debate about the stability of the bilingual advantage in attention-related skills in adulthood, what the current results highlight is that the differences observed during young and old adulthood between monolinguals and bilinguals are not observed during childhood. This, together with recent evidence showing larger differences in older than in younger participants (e.g., Gold et al., 2013), suggests a highly variable nature of the so-called bilingual advantage, which seems to be strongly dependent on a number of specific factors, among which the age of the samples should be carefully considered in future studies.

ACKNOWLEDGMENTS

This research was partially supported by grants CSD2008-00048, PSI2010-15133, PSI2011-23340, PSI2012-31448, and PSI2012-32123 from the Spanish Government, ERC-AdG-295362 from the European Research Council and PI2012-74 from the Basque Government. We wish to thank all the children and their families for kindly collaborating in this project and all the different schools for providing us with infrastructure. We are also very grateful to the research assistants who helped us in the data collection and to Margaret Gillon Dowens for her helpful comments.

REFERENCES

Anderson, P. (2002). Assessment and development of executive function (EF) during childhood. *Child Neuropsychol.* 8, 71–82. doi: 10.1076/chin.8.2.71.8724

Bialystok, E., Barac, R., Blaye, A., and Poulin-Dubois, D. (2010). Word mapping and executive functioning in young monolingual and bilingual children. *J. Cogn. Dev.* 11, 485–508. doi: 10.1080/15248372.2010.516420

Bialystok, E., Craik, F. I. M., Klein, R., and Viswanathan, M. (2004). Bilingualism, aging, and cognitive control: evidence from the Simon task. *Psychol. Aging* 19, 290–303. doi: 10.1037/0882-7974.19.2.290

Bialystok, E., Craik, F. I. M., and Luk, G. (2012). Bilingualism: consequences for mind and brain. *Trends Cogn. Sci.* 16, 240–250. doi: 10.1016/j.tics.2012.03.001

Bialystok, E., Craik, F., and Luk, G. (2008). Cognitive control and lexical access in younger and older bilinguals. *J. Exp. Psychol. Learn. Mem. Cogn.* 34, 859–873. doi: 10.1037/0278-7393.34.4.859

Colzato, L. S., Bajo, M. T., van den Wildenberg, W., Paolieri, D., Nieuwenhuis, S., La Heij, W., et al. (2008). How does bilingualism improve executive control? A comparison of active and reactive inhibition mechanisms. *J. Exp. Psychol. Learn. Mem. Cogn.* 34, 302–312. doi: 10.1037/0278-7393.34.2.302

Costa, A., Hernández, M., Costa-Faidella, J., and Sebastián-Gallés, N. (2009). On the bilingual advantage in conflict processing: now you see it, now you don't. *Cognition* 113, 135–149. doi: 10.1016/j.cognition.2009.08.001

Costa, A., Hernández, M., and Sebastián-Gallés, N. (2008). Bilingualism aids conflict resolution: evidence from the ANT task. *Cognition* 106, 59–86. doi: 10.1016/j.cognition.2006.12.013

Duñabeitia, J. A., Hernández, J. A., Antón, E., Macizo, P., Estévez, A., Fuentes, L. J., et al. (2014). The inhibitory advantage in bilnigual children revisited. *Exp. Psychol.* 61, 234–251. doi: 10.1027/1618-3169/a000243

Eriksen, B. A., and Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Percept. Psychophys.* 16, 143–149. doi: 10.3758/BF03203267

Fan, J., Flombaum, J. I., McCandliss, B. D., Thomas, K. M., and Posner, M. I. (2003). Cognitive and brain consequences of conflict. *Neuroimage* 18, 42–57. doi: 10.1006/nimg.2002.1319

Fan, J., Gu, X., Guise, K. G., Liu, X., Fossella, J., Wang, H., et al. (2009). Testing the behavioral interaction and integration of attentional networks. *Brain Cogn.* 70, 209–220. doi: 10.1016/j.bandc.2009.02.002

Fan, J., McCandliss, B. D., Fossella, J., Flombaum, J. I., and Posner, M. I. (2005). The activation of attentional networks. *Neuroimage* 26, 471–479. doi: 10.1016/j.neuroimage.2005.02.004

Fan, J., McCandliss, B. D., Sommer, T., Raz, A., and Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *J. Cogn. Neurosci.* 14, 340–347. doi: 10.1162/089892902317361886

Fan, J., and Posner, M. (2004). Human attentional networks. *Psychiatr. Prax.* 31(Suppl. 2), S210–S214. doi: 10.1055/s-2004-828484

Gathercole, V. C. M., Thomas, E. M., Kennedy, I., Prys, C., Young, N., Vinas Guasch, N., et al. (2014). Does language dominance affect cognitive performance in bilinguals? Lifespan evidence from preschoolers through older adults on card sorting, Simon, and metalinguistic tasks. *Front. Psychol.* 5:11. doi: 10.3389/fpsyg.2014.00011

Gold, B. T., Kim, C., Johnson, N. F., Kryscio, R. J., and Smith, C. D. (2013). Lifelong bilingualism maintains neural efficiency for cognitive control in aging. *J. Neurosci.* 33, 387–396. doi: 10.1523/JNEUROSCI.3837-12.2013

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism Lang. Cogn.* 1, 67–81. doi: 10.1017/S1366728998000133

Green, D. W., and Abulatebi, J. (2013). Language control in bilinguals: the adaptive control hypothesis. *J. Cogn. Psychol.* 25, 515–530. doi: 10.1080/20445911.2013.796377

Hernández, M., Costa, A., Fuentes, L. J., Vivas, A. B., and Sebastián-Gallés, N. (2010). The impact of bilingualism on the executive control and orienting networks of attention. *Bilingualism Lang. Cogn.* 13, 315–325. doi: 10.1017/S1366728909990010

Hilchey, M. D., and Klein, R. M. (2011). Are there bilingual advantages on nonlinguistic interference tasks? Implications for the plasticity of executive control processes. *Psychon. Bull. Rev.* 18, 625–658. doi: 10.3758/s13423-011-0116-7

Humphrey, A. D., and Valian, V. V. (2012) "Multilingualism and cognitive control: Simon and Flanker task performance in monolingual and multilingual young adults," in *Paper Presented at the 53rd Annual Meeting of the Psychonomic Society* (Minneapolis, MN).

Ishigami, Y., and Klein, R. M. (2010). Repeated measurement of the components of attention using two versions of the Attention Network Test (ANT): stability, isolability, robustness, and reliability. *J. Neurosci. Methods*, 190, 117–128. doi: 10.1016/j.jneumeth.2010.04.019

Kapa, L. L., and Colombo, J. (2013). Attentional control in early and later bilingual children. *Cogn. Dev.* 28, 233–246. doi: 10.1016/j.cogdev.2013.01.011

Kirk, N. W., Scott-Brown, K., and Kempe, V. (2013). "Do older Gaelic-English bilinguals show an advantage in inhibitory control?" in *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (Berlin).

Kousaie, S., and Phillips, N. A. (2012). Conflict monitoring and resolution: are two languages better than one? Evidence from reaction time and event-related brain potentials. *Brain Res.* 1146, 71–90. doi: 10.1016/j.brainres.2012.01.052

Kroll, J. F., and Bialystok, E. (2013). Understanding the consequences of bilingualism for language processing and cognition. *J. Cogn. Psychol. (Hove).* 25, 497–514. doi: 10.1080/20445911.2013.799170

Kroll, J. F., Bobb, S. C., Misra, M., and Guo, T. (2008). Language selection in bilingual speech: evidence for inhibitory processes. *Acta Psychol.* 128, 416–430. doi: 10.1016/j.actpsy.2008.02.001

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspect. Psychol. Sci.* 6, 299–312. doi: 10.1177/1745691611406925

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *J. Exp. Psychol. Gen.* 142, 573–603. doi: 10.1037/a0029146

Luk, G., Anderson, J. A. E., Craik, F. I. M., Grady, C., and Bialystok, E. (2010). Distinct neural correlates for two types of inhibition in bilinguals: response inhibition versus interference suppression. *Brain Cogn.* 74, 347–357. doi: 10.1016/j.bandc.2010.09.004

Mackie, M.-A., Van Dam, N. T., and Fan, J. (2013). Cognitive control and attentional functions. *Brain Cogn.* 82, 301–312. doi: 10.1016/j.bandc.2013.05.004

Morales, J., Gómez-Ariza, C. J., and Bajo, M. T. (2013). Dual mechanisms of cognitive control in bilinguals and monolinguals. *J. Cogn. Psychol.* 25, 531–546. doi: 10.1080/20445911.2013.807812

Morton, J. B., and Harper, S. N. (2007). What did Simon say? Revisiting the bilingual advantage. *Dev. Sci.* 10, 719–726. doi: 10.1111/j.1467-7687.2007.00623.x

Paap, K. R., and Greenberg, Z. I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cogn. Psychol.* 66, 232–258. doi: 10.1016/j.cogpsych.2012.12.002

Paap, K. R., and Liu, Y. (2014). Conflict resolution in sentence processing is the same for bilinguals and monolinguals: the role of confirmation bias in testing for bilingual advantages. *J. Neurolinguist.* 27, 50–74. doi: 10.1016/j.jneuroling.2013.09.002

Pelham, S. D., and Abrams, L. (2014). Cognitive advantages and disadvantages in early and late bilinguals. *J. Exp. Psychol. Learn. Mem. Cogn.* 40, 313–325. doi: 10.1037/a0035224

Posner, M. I. (1980). Orienting of attention. *Q. J. Exp. Psychol.* 32, 3–25. doi: 10.1080/00335558008248231

Prior, A., and MacWhinney, B. (2010). A bilingual advantage in task switching. *Bilingualism: Lang. Cogn.* 13, 253–262. doi: 10.1017/S1366728909990526

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* 16, 225–237. doi: 10.3758/PBR.16.2.225

Rueda, M. R., Fan, J., McCandliss, B. D., Halparin, J. D., Gruber, D. B., Lercari, L. P., et al. (2004). Development of attentional networks in childhood. *Neuropsychologia* 42, 1029–1040. doi: 10.1016/j.neuropsychologia.2003.12.012

Sawi, O., and Paap, K. (2013). "Test-retest reliability and convergent validity of measures of executive processing: evidence from the Simon, flanker, switching and antisaccade task," in *Poster presented at the meeting of the Cognitive Neuroscience Society* (San Francisco, CA).

Simon, J. R., and Rudell, A. P. (1967). Auditory S-R compatibility: the effect of an irrelevant cue on information processing. *J. Appl. Psychol.* 51, 300–304. doi: 10.1037/h0020586

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *J. Exp. Psychol.* 18, 643–662. doi: 10.1037/h0054651

Wang, H., and Fan, J. (2007). Human attentional networks: a connectionist model. *J. Cogn. Neurosci.* 19, 1678–1689. doi: 10.1162/jocn.2007.19.10.1678

Wu, Y. J., and Thierry, G. (2013). Fast modulation of executive function by language context in bilinguals. *J. Neurosci.* 33, 13533–13537. doi: 10.1523/JNEUROSCI.4760-12.2013

Yin, X., Zhao, L., Xu, J., Evans, A. C., Fan, L., Ge, H., et al. (2012). Anatomical substrates of the alerting, orienting and executive control components of attention: focus on the posterior parietal lobe. *PLoS ONE* 7:e50590. doi: 10.1371/journal.pone.0050590

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.

Received: 26 February 2014; accepted: 15 April 2014; published online: 07 May 2014.
Citation: Antón E, Duñabeitia JA, Estévez A, Hernández JA, Castillo A, Fuentes LJ, Davidson DJ and Carreiras M (2014) Is there a bilingual advantage in the ANT task? Evidence from children. Front. Psychol. ***5****:398. doi: 10.3389/fpsyg.2014.00398*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read, for greatest visibility

COLLABORATIVE PEER-REVIEW
Designed to be rigorous – yet also collaborative, fair and constructive

FAST PUBLICATION
Average 85 days from submission to publication (across all journals)

COPYRIGHT TO AUTHORS
No limit to article distribution and re-use

TRANSPARENT
Editors and reviewers acknowledged by name on published articles

SUPPORT
By our Swiss-based editorial team

IMPACT METRICS
Advanced metrics track your article's impact

GLOBAL SPREAD
5'100'000+ monthly article views and downloads

LOOP RESEARCH NETWORK
Our network increases readership for your article

Frontiers
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on