frontiers | Frontiers in Human Neuroscience

TYPE Perspective
PUBLISHED 11 November 2022
DOI 10.3389/fnhum.2022.1018708

Check for updates

OPEN ACCESS

EDITED BY
Antonio Benítez-Burraco,
University of Seville, Spain

REVIEWED BY
Nathan Oesch,
Western University, Canada
Dor Shilton,
Tel Aviv University, Israel

*CORRESPONDENCE
Shigeru Miyagawa
miyagawa@mit.edu

SPECIALTY SECTION
This article was submitted to
Speech and Language,
a section of the journal
Frontiers in Human Neuroscience

RECEIVED 13 August 2022
ACCEPTED 20 October 2022
PUBLISHED 11 November 2022

CITATION
Miyagawa S, Arévalo A and Nóbrega VA (2022) On the representation of hierarchical structure: Revisiting Darwin's musical protolanguage.
Front. Hum. Neurosci. 16:1018708.
doi: 10.3389/fnhum.2022.1018708

On the representation of hierarchical structure: Revisiting Darwin's musical protolanguage

Shigeru Miyagawa[1,2]*, Analía Arévalo[3] and Vitor A. Nóbrega[4]

[1]Department of Linguistics and Philosophy, Massachusetts Institute of Technology, Cambridge, MA, United States, [2]Institute of Biosciences, University of São Paulo, São Paulo, Brazil, [3]School of Medicine, University of São Paulo, São Paulo, Brazil, [4]Institute of Romance Studies, University of Hamburg, Hamburg, Germany

In this article, we address the tenability of Darwin's musical protolanguage, arguing that a more compelling evolutionary scenario is one where a prosodic protolanguage is taken to be the preliminary step to represent the hierarchy involved in linguistic structures within a linear auditory signal. We hypothesize that the establishment of a prosodic protolanguage results from an enhancement of a rhythmic system that transformed linear signals into speech prosody, which in turn can mark syntactic hierarchical relations. To develop this claim, we explore the role of prosodic cues on the parsing of syntactic structures, as well as neuroscientific evidence connecting the evolutionary development of music and linguistic capacities. Finally, we entertain the assumption that the capacity to generate hierarchical structure might have developed as part of tool-making in human prehistory, and hence was established prior to the enhancement of a prosodic protolinguistic system.

KEYWORDS

language, syntax, protolanguage, brain and language, birdsong, prosody

Introduction: Birdsong and language

Charles Darwin (1871, p. 55) noted that birdsong is the "nearest analogy to language." Just as songbirds have an instinct to sing, humans have an instinct to speak, and both species display a pre-mastery stage: subsongs in birds and babbling in humans (Aronov et al., 2008). These correlations led Darwin to conjecture that, prior to language, our ancestors were singing to communicate, what Fitch calls "musical protolanguage" (Fitch, 2005, 2006, 2010, 2013).

Recent studies show a surprising parallel between language and birdsong beyond simply sharing a pre-mastery stage (Yip, 2006, 2013; Bolhuis et al., 2010; Bolhuis and Everaert, 2013; Moorman and Bolhuis, 2013; Samuels, 2015; Miyagawa, 2017). In observing juvenile zebra finches (*Taeniopygia guttata*), Liu et al. (2004) identified two learning strategies. In "serial repetition," one syllable of the model is repeated and clearly articulated; in the motif strategy, the juvenile bird tries to imitate the tutor's vocal display in its entirety, and the articulation is noisy and imprecise. Similarly, O'Grady (2005) and others note that a human infant may adopt either the "analytic" style, which produces clearly articulated, one-word utterances, or the "*gestalt*" style, which produces large chunks of speech that are poorly articulated.

Regions in the forebrain controlling vocal production have been identified in humans as well as three independent lineages of songbirds (e.g., zebra finches; Pfenning et al., 2014). These regions display convergent specializations in the expression of 50–70 genes per brain region. Furthermore, in birds that do not sing (e.g., chickens, *Gallus gallus domesticus*) and a primate that does not have language (e.g., macaques; *Macaca fuscata*), no direct projection connects the vocal motor cortex to brainstem vocal motor neurons (Belyk and Brown, 2017; Nevue et al., 2020). Such observations endorse the assumption that language and birdsong share a common neurobiological substrate (Cahill et al., 2021) that would have allowed auditory-vocal learning, a capacity necessary for linguistic competence to emerge (Jarvis, 2019).

Taking Darwin's musical protolanguage as a starting point, we discuss the possible evolutionary scenario from a linear musical/rhythmic protolanguage to speech prosody that would develop into a full-fledged syntactic hierarchical system underlying language (de Rooij, 1975, 1976; Price et al., 1991; Schafer et al., 2000; Richards, 2010, 2016, 2017; Speer et al., 2011; Langus et al., 2012; a.o.). To develop this claim, we explore the role of prosodic cues on the parsing of syntactic structures, as well as neuroscientific evidence connecting the evolutionary development of musical and linguistic capacities. Finally, we entertain the assumption that the capacity to generate hierarchical structure might have developed as part of tool-making prior to language.

Musical protolanguage

Like birdsong, Darwin (1871) assumed that the earliest musical protolanguage did not contain any propositional meaning. Birds sing to convey intention, typically the desire to mate (Marler, 1998, 2000; Berwick et al., 2011; Berwick et al., 2013; Bowling and Fitch, 2015). Darwin (1871, p. 56–57) conjectured that the musical protolanguage was for "charming the opposite sex." Given the lack of meaning, this musical protolanguage by itself could not have developed into human language. Darwin suggested that our ancestors began to interweave gestures and sound imitations of other animals as precursors to words in order to insert meaning into the musical sequences.

In the same vein, but with more knowledge about human language than what was available to Darwin, Fitch (2005, 2010, 2013) suggests that for the musical protolanguage to have transformed into language, a second stage must have added "a fully propositional and intentional semantics" (2005:220; see also Fitch, 2004). Fitch suggested there was an integration of existing systems: the musical protolanguage and the propositional system. More specifically, Fitch's version of a musical protolanguage expands Darwin's original formulation by offering an account of how an intentional semantics — as opposed to lexical semantics— was assigned to melodic strings, as well as how modern humans developed advanced vocal control and learning; a major obstacle for a cohesive explanation on the phylogenetic history of a linguistic capacity. In this article, we argue instead that complex vocal control, which paved the way for singing and rhythmic utterances, might have enhanced a parsing mechanism for syntactic constituency, hence for the identification of hierarchic structures, by means of prosodic cues (e.g., pauses, prominence, nuclear stress, etc.). Fitch (2010, p. 499) also refers to his model as a "prosodic" protolanguage, which "[. . .] consisted of sung syllables, but *not* of notes that could be arranged in a scale, nor produced with a steady rhythm" (see also Fitch, 2006). His prosodic protolanguage model, however, focuses on the evolutionary development of prosodic units rather than on the impact of prosodic cues in the identification of syntactic hierarchical structure, as we are proposing.

Miyagawa et al. (2013, 2014) and Miyagawa (2017) note that components of human language existed long before language emerged.[1] These components became integrated in recent evolutionary time, perhaps around 300–200 thousand years ago (kya) (Tattersall, 2008, 2010, 2012, 2016; Huybregts, 2017), to give form to language as we know it today. This integration of the musical protolanguage with the propositional component, as envisaged by Fitch, would have been a very complex process. Human language is associated with the core syntactic component, which generates structured phrases, and the interfaces to which the structured phrases are sent: the phonological form (PF), which connects to the sensory-motor system, and is responsible for the externalization of the structured phrases; and the logical form (LF), which connects to the conceptual-intentional system, assigning an interpretation to the structure (Chomsky, 1995, 2000; see **Figure 1**).

We argue that a prosodic protolanguage, resulting from complex vocal control —fundamental for singing and rhythmic vocal displays—, would have been part of the PF component, enabling externalization of the core syntactic component. For this to happen, it developed the capacity to represent hierarchy within a linear signal. This proposal, when compared to Fitch's, has the benefit of being more easily tested, since we can assess whether the absence of prosodic cues lead to divergent/unexpected parsing strategies or makes syntactic interpretation difficult.[2] By pulling together research from

[1] See also Fitch (2002) and Hauser et al. (2002).

[2] For example, Sandler et al. (2011) stress the role of prosodic cues in the development of syntactic complexity when analyzing the

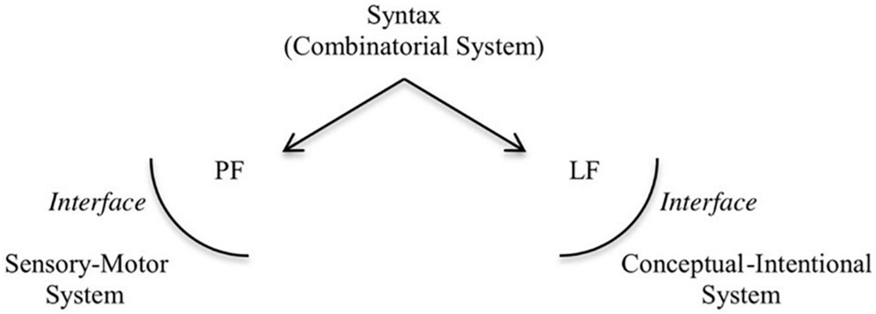


FIGURE 1
The architecture of the human language faculty.

neuroscience, primatology, and linguistics, we develop in this article a reasonably coherent picture of how hierarchy might have emerged in speech.[3]

One region that has been implicated in the creation of hierarchical relations is Broca's area, specifically, the pars opercularis, or Brodmann area 44 (BA44) (Friederici et al., 2006; Friederici, 2009; Friederici et al., 2012; Kemmerer, 2012, 2015, 2021; Zaccarella and Friederici, 2015a,b,c). Studies have also explored the evolution of this region in humans and its homologs in other species, such as the great apes. These studies suggest that human BA44 is proportionately much larger than its homolog in other species (compared with the entire brain or specific regions like the entire frontal cortex; see Schenker et al., 2010; Smaers et al., 2017; Donahue et al., 2018), and that left BA44 in humans may have greater neuropil volume, suggesting greater space for local and inter-regional connectivity (Palomero-Gallagher and Zilles, 2019; Changeaux et al., 2021). We explore the idea that if the musical protolanguage played a role in the evolution of language by transforming into what we call speech prosody, as Darwin originally suggested, it may have involved BA44 and its critical connections to other regions.[4]

Prosody

Words in language are uttered in a linear fashion. The words are not simply linearly ordered but are also hierarchically organized, and this hierarchy comprises the essential component for associating meaning to the expression. The hierarchy itself is an abstract representation, and is commonly communicated by prosody, as a layer of supra-segmental phonological information on top of the string of words (e.g., Selkirk, 1986; Jackendoff, 1997; Büring, 2013). There are two types of prosody: emotional and linguistic. Emotional prosody signals the speaker's emotional state or the emotional content of the expression, while linguistic prosody signals syntactic structure and thematic relations.[5] Here we will focus on the latter. We give three examples of such prosody: (i) pauses, which mark clausal structure, (ii) relative prominence assigned to units within a noun phrase, and (iii) nuclear stress, which is assigned within a verb phrase.

Pause

The following shows how pause, or major prosodic constituents, can be placed within a sentence (from Büring, 2013, p. 865).

development of the Al-Sayyid Bedouin emergent sign language. In this language, rhythmic and facial cues are directly aligned at constituent boundaries. The importance of prosodic cues in the development of syntactic constituency can further be tested in other nascent linguistic system that lack any previous linguistic bias, such as the Cena rural sign language in Northeast Brazil (Almeida-Silva and Nevins, 2020).

[3] It is relevant to point out that Benítez-Burraco and Elvira-García (2022) reach similar conclusions by exploring the role of self-domestication in the evolutionary development of speech prosody. In their view prosody, which is argued to have been affected by human self-domestication, might have favored syntactic complexification through a series of bootstrapping effects.

[4] Katz and Pesetsky (2011) and Roberts (2012) show that both music and language employ a parallel computation for hierarchical structure building. We acknowledge that the cognitive mechanisms underlying hierarchical structure in both music and language might have had a common ancestry, as will be explored later (see also Jackendoff, 2009; Boeckx and Fujita, 2014; Fitch and Martins, 2014; Asano and Boeckx, 2015; Asano, 2021; Asano et al., 2022).

[5] Prosody often marks structure in neutral focus. If there is narrow focus in way of stress for emphasis, prosody does not necessarily mirror the structure of the expression (Ladd, 2008). Some languages, however, seem to involve a different pattern. Shanghainese and some Bantu languages display a mismatch between prosody and syntactic structure in neutral focus (e.g., Zubizarreta, 2009; Han et al., 2013). This linguistic variability with respect to prosodically marked neutral focus led some linguists to suggest that prosody may not have a faithful one-to-one mapping from syntax, being responsible for mapping only certain syntactic domains (Selkirk, 2009, 2011).

(1) when Roger left the house became irrelevant.

(a) when Roger left [PAUSE] the house became irrelevant

(b) when Roger left the house [PAUSE] became irrelevant

(1) shows how pauses indicate structural boundaries. The silent intervals in (1a) and (1b) signal the end of a subordinate clause, with the varying positions leading to different interpretations.[6]

Prominence: Noun phrase

Speakers can tell which syllable is prominent in an utterance. Prominence can often be measured by duration, intensity, fundamental frequency (pitch) and other acoustic measures. Prominent syllables tend to be longer and louder. So, a syllable (along with the word that contains it) is perceived as prominent if it is in the location of the local maximum in the fundamental frequency curve. Conversely, it is perceived as less prominent if it is in the location of the local minimum in the fundamental frequency curve (see Büring, 2013, and references therein). In English, very roughly, the last syllable/word in a constituent receives relative prominence (e.g., Selkirk, 1986). The following is modeled on similar examples from Büring (2013).

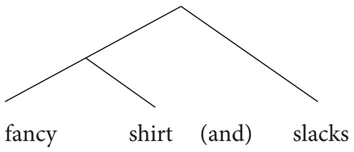
(2) a.				(*)
		(*)		(*)
	(*)	(*)		(*)
	fancy	shirt	and	slacks
				(*)
	(*)	(*)		(*)
b.	tie,	shirt	and	slacks

The number of asterisks indicates relative prominence. In (2a), *fancy* and *shirt* differ in prominence, with *shirt* receiving more prominence. This indicates that *shirt* is at the right edge of the phrase that also contains *fancy*. The third word, *slacks*, receives more prominence than *shirt*, indicating that it is at the right edge of another phrase.

(3) [[fancy shirt] and slacks]

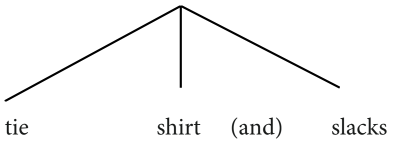
This is a hierarchical relation, with *fancy shirt* in the lower tier of the hierarchy.

(4)



In (2b), no distinction exists between *tie* and *shirt*, so these words do not constitute a phrase. The relative prominence of the last word, *slacks*, shows that this word is on the right edge of the entire phrase: [*tie shirt and slacks*].

(5)



Prominence: Nuclear stress rule

Within a verb phrase of a sentence with neutral focus, a rhythmically prominent stress falls on a particular constituent, called Nuclear Stress (NS) (Chomsky and Halle, 1968; see also Zubizarreta, 1998; Reinhart, 2006). The NS in the example below falls on *book*, the final element in the verb phrase (and the sentence).

(6) Mary read a book.

There is general recognition that syntactic structure plays a crucial role in the assignment of NS (e.g., Chomsky, 1971; Jackendoff, 1972; Cinque, 1993; Selkirk, 1995; Kahnemuyipour, 2004, 2009; Reinhart, 2006; Truckenbrodt, 2006; Kratzer and Selkirk, 2007; Féry, 2011). It appears at first that the NS is assigned to the last element in the sentence. This would be a linearly based analysis of NS. A key observation for the structurally based NS assignment is that in a language such as German, where the object precedes the verb, the NS falls not on the final element, but on the object, just as in English.

(7) *Hans hat ein **Buch** gelesen.*
Hans has a book read
"Hans has read a book."

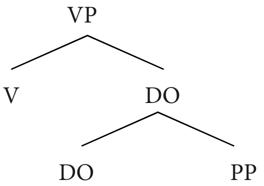
[6] Yip (2013, p. 191) indicates that a "motif" could be roughly equated to a phrase, "in its tendency to be surrounded by 'pauses"'. Such category in birdsong plays a crucial role during ontogeny, since infants first begin copying small chunks of the target song. Williams and Staples (1992) show that the chunk boundaries produced by infants correlate with the silent interval delimited by the pauses circumscribing a motif, suggesting that similar acoustic cues assist on the identification of the internal structure of a song, facilitating its segmentation — a strategy that is parallel to the prosodic bootstraps in language acquisition (Yip, 2013; Mol et al., 2017). Song segmentation, however, seems to be circumscribed to the identification of which note strings might comprise a motif, and which are the linear organizations of motifs into a complete song. Birdsong involves a finite-state mechanism to combine notes into motifs, and motifs into songs (Berwick et al., 2011, Berwick et al., 2012). A finite-state mechanism resorts to strictly sequential steps (linear probability), hence lacks hierarchical organization. The latter is only available in combinatorial systems that demand a more powerful working memory, such as context-free or context-sensitive systems (Joshi, 1985), which was not observed in songbirds.

In either order, English or German, the verb and the object are in the verb phrase: [$_{\text{VP}}$ Verb OBJ]. There is an assumption that the verb must vacate the verb phrase and move to a higher position, leaving, in this case, only the object: [$_{\text{VP}}$ __ OBJ]. Is it always the object that is assigned the NS? The example below shows that it is not.

(8) Mary read a book about the **moon**.

The NS in (8) falls on *moon* within the prepositional phrase that follows the object. This indicates that the NS is assigned to the highest element in the verb phrase (Kahnemuyipour, 2004, 2009; Kratzer and Selkirk, 2007).

(9)



The NS assignment is not dependent on linear order, but strictly on hierarchical structure. In this way, speech prosody marks hierarchy.[7]

Music and prosody

Some evolutionary theories contend that music and language have a common progenitor that gave rise to an early communication system (Brown, 2001; Mithen, 2005). Both human speech and music contain prosody, which in turn contains melody (intonation) and rhythm (stress and timing) (Nooteboom, 1997; see also Yip, 2013). Music and prosody have been shown to recruit overlapping neural regions, supporting Darwin's original idea and the evolutionary theories that it spawned (Peretz et al., 1994; Patel, 2008, 2012). Some have suggested that language and music are on a continuum, without a sharp line of demarcation (Jackendoff, 2009; Patel, 2010; Koelsch, 2012). Early in life, infant-directed speech (IDS), or "motherese" (Gleitman et al., 1984; Bates et al., 1995; de Boysson-Bardies, 1999) seems to imitate song, and infants show overlapping neural activity to IDS and instrumental music (Kotilahti et al., 2010).

In studies of amusia without aphasia, Patel et al. (1998) observed that prosodic and musical discrimination were preserved or affected together, suggesting that the perception of prosody and musical contour share overlapping cognitive and neural resources.[8] Furthermore, studies showing that individuals with a congenital deficit in music perception typically also exhibit deficits in perception of pitch in language (Peretz, 1993; Liu et al., 2010; Nan et al., 2010; Tillmann et al., 2011).

Over the last several decades, melodic intonation therapy (MIT) has been used to improve language production in patients with aphasia. Often, these patients have global aphasia and respond poorly to other forms of classical therapies. Patients who benefit from MIT may be activating remaining frontoparietal networks critical to language, music and motor processing (Sparks et al., 1974; Leonardi et al., 2017).

According to Hausen et al. (2013), studies using fMRI have shown that music and language recruit overlapping neural regions, including superior, anterior and posterior temporal, parietal, and inferior frontal areas (Koelsch et al., 2002; Tillmann et al., 2003; Brown and Martinez, 2007; Rauschecker and Scott, 2009; Schön et al., 2010; Abrams et al., 2011; Rogalsky et al., 2011).

While music and prosody are largely processed in the right hemisphere of the brain (Weintraub et al., 1981; Bradvik et al., 1991), hierarchy is associated with left Broca's area (BA44) (Friederici et al., 2006; Friederici, 2009; Friederici et al., 2012; Zaccarella and Friederici, 2015a,b,c). Meyer et al. (2002) showed that speech normally recruits both hemispheres, while prosodic speech without any segmental information activates mostly the right hemisphere. Speech processing streams connect the hemispheres via the posterior portion of the corpus callosum. As evidence of this, syntax-prosody mismatches in an ERP paradigm did not elicit an anterior negativity in patients with lesions to the posterior third of the corpus callosum (vs. patients with lesions to the anterior two-thirds of the corpus callosum and controls) (Sammler et al., 2010).

Stone tools: Source of hierarchy?

If BA44 is a critical piece of the puzzle when it comes to generating hierarchy, then presumably the original musical protolanguage would have undergone enhancement by connecting to this region to produce

[7] Further prosodic phenomena responsible for marking constituent boundaries are (i) stress prominence in English, which normally falls on the rightmost constituent within a phrase (e.g., [[*A sènator* [*from Chicágo*]] [*wòn* [*the làst eléction*]]] (Chomsky and Halle, 1968 *apud*<ref> Selkirk, 2011, p. 435), (ii) *liason* in French, i.e., maintenance of a word-final consonant before a vowel, [[Le peti**t** **â**ne] [le suivait]] "The little donkey followed him" vs. [[*Le peti*] [[*aime*] [*le Guignol*]] "The little one loves the puppet theater" (Selkirk, 1974 *apud* Selkirk, 2011, p. 435– 436). Several additional phenomena can be found in Selkirk (2011). In sign languages, non-manual markers, such as head position and facial expression, serve the role of prosodic cues, and are equally relevant for syntactic parsing involved in topicalization, relative clauses, and *wh*-constructions (see Baker and Padden, 1978; Liddell, 1978, 1980; Neidle et al., 2000, for American Sign Language).

[8] Earlier studies have reported a dissociation between the processing of language and music (Marin, 1989; Peretz and Morais, 1989, 1993; Sergent, 1993). See Patel (2012) for comments on this apparent dissociation.

speech prosody. Under this view, the capacity to generate hierarchical structures existed prior to the enhancement. If so, how did the capacity to generate hierarchical structure develop? One view is that hierarchical cognition developed as part of tool-making, as initially suggested by Lashley (1951), and recently expanded by Fitch and Martins (2014), Asano and Boeckx (2015), and Asano (2021). This idea, which is controversial (Putt et al., 2017), was primarily developed by Greenfield's grammars of action (Greenfield, 1991, 1998). From their studies with non-human primates, Greenfield and colleagues suggested three general "grammatical" strategies: pairing strategy, pot strategy, and subassembly strategy; this last one, subassembly, requires hierarchical organization of information. They observed that while non-human primates could engage in the first two strategies, only humans are capable of the third strategy, suggesting hierarchical organization is an exclusively human trait.

A large body of work has applied this general approach to stone tools, with the assumption that higher cognitive functions in modern humans are linked with the evolution of motor control (Lieberman, 2006; see also Holloway, 1969; Wynn, 1991; Fitch and Martins, 2014). Stone tools are made from flake units, which are combined to form assemblies, and these assemblies make up the tool's higher-order architecture (Miller et al., 1960). Earlier (i.e., Pleistocene era) tools do not evidence this kind of hierarchical structure. Moore (2010) argues that it appeared in late Middle Pleistocene, around 270 kya, when the Mousterian style of tool-making appeared with the Neanderthals; however, rudimentary hierarchical cognition may have supported tool-making much earlier, approximately 800 kya or earlier, during the Acheulean phase (Moore, 2010; Stout and Hecht, 2014; Gaucherel and Noûs, 2020).[9] If true, the capacity for hierarchical cognition existed long before human language emerged. If so, this baseline would have allowed the musical protolanguage to evolve and give rise to speech prosody. Additional support for these ideas comes from imaging studies showing overlapping activations for language and tool use tasks (Stout et al., 2008; Higuchi et al., 2009; Stout and Chaminade, 2012; Osiurak et al., 2021).[10]

What came first?

In this article, we traced our arguments beginning with Darwin's original suggestion that "[. . .] musical cries by articulate sounds may have given rise to words expressive of various complex emotions" (Darwin, 1871; see also Oesch, 2020). This statement implies the following sequence of emerging functions: isolated melodic cries, then complex vocalizations (with increasing articulatory refinement), then simple linguistic utterances, followed by increasingly complex language containing words capable of conveying emotions. A parallel theory suggests music and language may have evolved simultaneously on a spectrum (Morley, 2013; Oesch, 2019). This last theory gains strength in the fact that fossil records — the only direct source of information on this matter— are inherently limited, which currently precludes us from determining causality.

Thus, given these limitations, an equally plausible proposal would be the reverse: that speech in fact preceded music. Here we list a few arguments that make this possibility less convincing. As mentioned above, studies have revealed an expansion of several cortical regions (e.g., BA44, auditory-vocal cortical regions) as well as sensorimotor connectivity in humans relative to non-human primates, which is thought to have permitted the enhancement of critical components of language, including vocal working memory and vocal repertoire size (Schenker et al., 2010; Smaers et al., 2017; Aboitiz, 2018; Donahue et al., 2018; Ardesch et al., 2019; Palomero-Gallagher and Zilles, 2019; Changeaux et al., 2021). Compared with non-human primates and other species known to engage in "cooperative vocal turn-taking,"[11] humans arguably have the most complex language, at least in terms of vocabulary size and internal structure. Thus, the work in comparative neuroanatomy and connectivity would suggest that language, at least in its most evolved, modern state, would not have emerged earlier than musical abilities.

Although archeologists have suggested that the fine motor control required for modern-day vocalizations may have been present in *Homo heidelbergensis* as early as 5–800,000 years ago (MacLarnon and Hewitt, 1999; Martinez et al., 2013; Oesch, 2019), some forms of musical

[9] More specifically, Moore (2010) shows that hierarchical flaking is necessary for stone tool types that demand multiple preparatory steps prior to a flake removal, such as Acheulean bifaces and the Levallois method. The production of Oldowan choppers, differently from bifaces and the Levallois' core preparation, only requires the extraction of high mass from the core, lacking preparatory flaking (see also Stout, 2011; Stout et al., 2018, for similar conclusions).

[10] It is relevant to point out that vocal learning and vocal control evolved independently from language (Jarvis, 2004, Jarvis, 2019), hence prior to syntactic structuring. We also find suggestive evidence that hierarchy was presumably co-opted from the abilities involved in the motor actions of stone tool-making (see Fitch and Martins, 2014; Asano and Boeckx, 2015; Asano, 2021). With this timeline in mind, we can entertain an evolutionary scenario where complex vocal control, roughly understood as an embryonic stage of prosodic cues, might have enhanced the representation of hierarchic structure in the expressive utterances of early human, gradually leading to present-day syntax. In this scenario, we can say that prosody and syntactic structuring co-evolved.

[11] According to Oesch (2019), these are a rare type of vocalization that bridges the gap between animal calls and human speech.

expressions, such as drumming or marking a beat (e.g., beat entrainment), do not require any vocalizations at all. So, in line with the above arguments, the evolutionary record would suggest that the biological substrates and mechanisms required for music production would have been in place before those for the most advanced forms of language. However, several authors have argued that beat entrainment requires fine motor control, including vocal control (see Patel, 2021; Shilton, 2022).[12] With this in mind, we can speculate that until fine motor control and vocalization systems to support musical as well as linguistic communication emerged in early hominins, it is very likely that gestures might have played an even more prominent role in communication.

So, if the fossil record is limited, what can other lines of research contribute to elucidating these questions? One hope lies in modern neuroscientific research. As our technologies advance at unprecedented rates, well-designed studies using connectivity, electrophysiology, electrocorticography, and coherence should test musical and language processing in humans as well as other species. As we become progressively closer to understanding the real time processes involved in different forms of musical and linguistic processing, we can further our understanding of how evolutionarily more recent structures may have supported such processes, thus providing evidence for or against theories tracing the sequential or parallel emergence of these skills.

Concluding remarks

Darwin's musical protolanguage, if it existed, must have undergone many critical changes before it became modern-day language. One crucial step would have been tapping into the ability to produce hierarchical structure, which is only present in human language. We suggest that this step involved enhancement of the musical system to transform it to speech prosody, which can mark hierarchical relations. Other steps were needed for the hierarchical structure marked by prosody to link up with a fully propositional intentional semantics. But it is a crucial step, as we can see by the pervasive nature of hierarchical structure in human language.

Data availability statement

The original contributions presented in this study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

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

Funding

SM and AA's research was funded by the São Paulo Research Foundation (FAPESP) (grant no. 2018/18900-1), research project "Innovations in Human and Non-Human Animal Communities," from which the results here presented were part. VN's research was funded by the German Academic Exchange Service (DAAD) (grant no. 57604641).

Acknowledgments

We thank the two reviewers and the associate editor for numerous helpful suggestions. We also thank Danfeng Wu and Bill Idsardi for comments on an earlier draft.

Conflict of interest

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

Publisher's note

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

[12] We thank one of the reviewers for suggesting us this point.

References

Aboitiz, F. (2018). A brain for speech. Evolutionary continuity in primate and human auditory-vocal processing. *Front. Neurosci.* 12:174. doi: 10.3389/fnins.2018.00174

Abrams, D. A., Bhatara, A., Ryali, S., Balaban, E., Levitin, D. J., and Menon, V. (2011). Decoding temporal structure in music and speech relies on shared brain resources but elicits different fine-scale spatial patterns. *Cereb. Cortex* 21, 1507–1518. doi: 10.1093/cercor/bhq198

Almeida-Silva, A., and Nevins, A. (2020). Observações sobre a estrutura linguística da Cena: A língua de sinais emergente da Várzea Queimada (Piauí. Brasil). *Rev. Ling. Ensino* 23, 1029–1053.

Ardesch, D. J., Scholtens, L. H., Li, L., and van den Heuvel, M. P. (2019). Evolutionary expansion of connectivity between multimodal association areas in the human brain compared with chimpanzees. *Proc. Natl. Acad. Sci. U.S.A.* 116, 7101–7106. doi: 10.1073/pnas.1818512116

Aronov, D., Andalman, A., and Dee, M. (2008). A specialized forebrain circuit for vocal babbling in the juvenile songbird. *Science* 320, 630–634. doi: 10.1126/science.1155140

Asano, R. (2021). The evolution of hierarchical structure building capacity for language and music: A bottom-up perspective. *Primates* 63, 417–428. doi: 10.1007/s10329-021-00905-x

Asano, R., and Boeckx, C. (2015). Syntax in language and music: What is the right level of comparison? *Front. Psychol.* 6:942. doi: 10.3389/fpsyg.2015.00942

Asano, R., Boeckx, C., and Fujita, K. (2022). Moving beyond domain-specific versus domain-general options in cognitive neuroscience. *Cortex* 154, 259–268. doi: 10.1016/j.cortex.2022.05.004

Baker, C., and Padden, C. (1978). "Focusing on the nonmnanual components of ASL," in *Understanding language through sign language research*, ed. P. Siple (New York, NY: Academic Press), 27–57.

Bates, E., Dale, P. S., and Thal, D. (1995). "Individual differences and their implications for theories of language development," in *The handbook of child language*, eds P. Fletcher and B. MacWhinney (Oxford: Blackwell Publishers).

Belyk, M., and Brown, S. (2017). The origins of the vocal brain in humans. *Neurosci. Biobehav. Rev.* 77, 177–193. doi: 10.1016/j.neubiorev.2017.03.014

Benítez-Burraco, A., and Elvira-García, W. (2022). Human self-domestication and the evolution of prosody. *PsyArXiv* [Preprint]. doi: 10.31234/osf.io/8uzht

Berwick, R. C., Beckers, G. J. L., Okanoya, K., and Bolhuis, J. J. (2012). A bird's eye view of human language evolution. *Front. Evol. Neurosci.* 4:5. doi: 10.3389/fnevo.2012.00005

Berwick, R. C., Friederici, A. D., Chomsky, N., and Bolhuis, J. J. (2013). Evolution, brain, and the nature of language. *Trends Cogn. Sci.* 17, 89–98. doi: 10.1016/j.tics.2012.12.002

Berwick, R. C., Okanoya, K., Beckers, G. J. L., and Bolhuis, J. J. (2011). Songs to syntax: The linguistics of birdsong. *Trends Cogn. Sci.* 15, 113–121. doi: 10.1016/j.tics.2011.01.002

Boeckx, C., and Fujita, K. (2014). Syntax, action, comparative cognitive science, and Darwinian thinking. *Front. Psychol.* 5:627. doi: 10.3389/fpsyg.2014.00627

Bolhuis, J. J., and Everaert, M. (2013). *Birdsong, speech and language. Exploring the evolution of mind and brain*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9322.001.0001

Bolhuis, J. J., Okanoya, K., and Scharff, C. (2010). Twitter evolution: Converging mechanisms in birdsong and human speech. *Nat. Rev. Neurosci.* 11, 747–759. doi: 10.1038/nrn2931

Bowling, D. L., and Fitch, W. T. (2015). Do animal communication systems have phonemes? *Trends Cogn. Sci.* 19, 555–557. doi: 10.1016/j.tics.2015.08.011

Bradvik, B., Dravins, C., Holtas, S., Rosen, I., Ryding, E., and Ingvar, D. H. (1991). Disturbances of speech prosody following right hemisphere infarcts. *Acta Neurol. Scand.* 54, 114–126. doi: 10.1111/j.1600-0404.1991.tb04919.x

Brown, S. (2001). "The 'musilanguage' model of music evolution," in *The origins of music*, eds N. L. Wallin, B. Merker, and S. Brown (Cambridge, MA: MIT Press), 271–300. doi: 10.7551/mitpress/5190.003.0022

Brown, S., and Martinez, N. J. (2007). Activation of premotor vocal areas during musical discrimination. *Brain Cogn.* 63, 59–69. doi: 10.1016/j.bandc.2006.08.006

Büring, D. (2013). "Syntax, information structure, and prosody," in *The Cambridge handbook of generative syntax*, ed. M. den Dikken (Cambridge: Cambridge University Press), 860–895. doi: 10.1017/CBO9780511804571.029

Cahill, J. A., Armstrong, J., Deran, A., Khoury, C. J., Paten, B., Haussler, D., et al. (2021). Positive selection in noncoding genomic regions of vocal learning birds is associated with genes implicated in vocal learning and speech functions in humans. *Genome Res.* 31, 1–15. doi: 10.1101/gr.275989.121

Changeaux, J. P., Goulas, A., and Hilgetag, C. C. (2021). A connectomic hypothesis for the hominization of the brain. *Cereb. Cortex* 31, 2425–2449. doi: 10.1093/cercor/bhaa365

Chomsky, N. (1971). "Deep structure, surface structure, and semantic interpretation," in *Semantics: An interdisciplinary reader in philosophy, linguistics, and psychology*, eds D. Steinberg and L. Jakobovits (Cambridge: Cambridge University Press).

Chomsky, N. (1995). *The minimalist program*. Cambridge, MA: MIT Press.

Chomsky, N. (2000). "Minimalist inquiries: The framework," in *Step by step: Essays on minimalist syntax in honor of Howard Lasnik*, eds R. Martin, D. Michaels, and J. Uriagereka (Cambridge, MA: MIT Press), 89–155.

Chomsky, N., and Halle, M. (1968). *The sound pattern of english*. New York, NY: Harper and Row.

Cinque, G. (1993). A null theory of phrase and compound stress. *Linguist. Inq.* 24, 239–297.

Darwin, C. (1871). *The descent of man, and selection in relation to sex*. London: John Murray.

de Boysson-Bardies, B. (1999). *How language comes to children, from birth to two years*, trans. M. DeBevoise. Cambridge, MA: MIT Press.

de Rooij, J. J. (1975). Prosody and the perception of syntactic boundaries. *IPO Annu. Prog. Rep.* 10, 36–39.

de Rooij, J. J. (1976). Perception of prosodic boundaries. *IPO Annu. Prog. Rep.* 11, 20–24.

Donahue, C. J., Glasser, M. F., Preuss, T. M., Rilling, J. K., and Van Essen, D. C. (2018). Quantitative assessment of prefrontal cortex in humans relative to nonhuman primates. *Proc. Natl. Acad. Sci. U.S.A.* 115, E5183–E5192. doi: 10.1073/pnas.1721653115

Féry, C. (2011). German sentence accents and embedded prosodic phrases. *Lingua* 121, 1906–1922. doi: 10.1016/j.lingua.2011.07.005

Fitch, W. T. (2002). The evolution of language comes of age. *Trends Cogn. Sci.* 6, 278–279. doi: 10.1016/S1364-6613(02)01925-3

Fitch, W. T. (2004). "Kin selection and "Mother Tongues": A neglected component in language evolution," in *Evolution of communication systems: A comparative approach*, eds D. K. Oller and U. Griebel (Cambridge, MA: MIT Press), 275–296.

Fitch, W. T. (2005). The evolution of language: A comparative review. *Biol. Philos.* 20, 193–230. doi: 10.1007/s10539-005-5597-1

Fitch, W. T. (2006). The biology and evolution of music: A comparative perspective. *Cognition* 100, 173–215. doi: 10.1016/j.cognition.2005.11.009

Fitch, W. T. (2010). *The evolution of language*. Cambridge: Cambridge University Press.

Fitch, W. T. (2013). "Musical protolanguage: Darwin's theory of language evolution revisited," in *Birdsong, speech, and language: Exploring the evolution of mind and brain*, eds J. J. Bolhuis and M. Everaert (Cambridge, MA: MIT Press), 489–503. doi: 10.7551/mitpress/9322.003.0032

Fitch, W. T., and Martins, M. D. (2014). Hierarchical processing in music, language, and action: Lashley revisited. *Ann. N. Y. Acad. Sci.* 1316, 87–104. doi: 10.1111/nyas.12406

Friederici, A. D. (2009). Pathways to language: Fiber tracts in the human brain. *Trends Cogn. Sci.* 13, 175–181. doi: 10.1016/j.tics.2009.01.001

Friederici, A. D., Bahlmann, J., Heim, S., Schubotz, R. I., and Anwander, A. (2006). The brain differentiates human and non-human grammars: Functional localization and structural connectivity. *Proc. Natl. Acad. Sci. U.S.A.* 103, 2458–2463. doi: 10.1073/pnas.0509389103

Friederici, A. D., Oberecker, R., and Brauer, J. (2012). Neurophysiological preconditions of syntax acquisition. *Psychol. Res.* 76, 204–211. doi: 10.1007/s00426-011-0357-0

Gaucherel, C., and Noûs, C. (2020). *Platforms of palaeolithic knappers reveal complex linguistic abilities*. PCI Archeo. doi: 10.31233/osf.io/wn5za

Gleitman, L. R., Newport, E. L., and Gleitman, H. (1984). The current status of the motherese hypothesis. *J. Child Lang.* 11, 43–79. doi: 10.1017/S0305000900005584

Greenfield, P. M. (1991). Language, tools and brain: The ontogeny and phylogeny of hierarchically organized sequential behavior. *Behav. Brain Sci.* 14, 531–595. doi: 10.1017/S0140525X00071235

Greenfield, P. M. (1998). Language, tools, and brain revisited. *Behav. Brain Sci.* 21, 159–163. doi: 10.1017/S0140525X98230962

Han, W., Arppe, A., and Newman, J. (2013). Topic marking in a Shanghainese corpus: From observation to prediction. *Corpus Linguist. Linguist. Theory* 13, 291–319. doi: 10.1515/cllt-2013-0014

Hausen, M., Torppa, R., Salmela, V. R., Vainio, M., and Särkämö, T. (2013). Music and speech prosody: A common rhythm. *Front. Psychol.* 4:566. doi: 10.3389/fpsyg.2013.00566

Hauser, M. C., Chomsky, N., and Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science* 298, 1569–1579. doi: 10.1126/science.298.5598.1569

Higuchi, S., Chaminade, T., Imamizu, H., and Kawato, M. (2009). Shared neural correlates for language and tool use in Broca's area. *Neuroreport* 20, 1376–1381. doi: 10.1097/WNR.0b013e3283315570

Holloway, R. L. (1969). Culture: A human domain. *Curr. Anthropol.* 10, 395–412. doi: 10.1086/204018

Huybregts, M. A. C. R. (2017). Phonemic clicks and the mapping asymmetry: How language emerged and speech developed. *Neurosci. Biobehav. Rev.* 81, 279–294. doi: 10.1016/j.neubiorev.2017.01.041

Jackendoff, R. (1972). *Semantic interpretation in generative grammar*. Cambridge, MA: MIT Press.

Jackendoff, R. (1997). *The architecture of the language faculty*. Cambridge, MA: MIT Press.

Jackendoff, R. (2009). Parallels and non-parallels between language and music. *Music Percept.* 26, 195–204. doi: 10.1525/mp.2009.26.3.195

Jarvis, E. D. (2004). Learned birdsong and the neurobiology of human language. *Ann. N. Y. Acad. Sci.* 1016, 749–777. doi: 10.1196/annals.1298.038

Jarvis, E. D. (2019). Evolution of vocal learning and spoken language. *Science* 366, 50–54. doi: 10.1126/science.aax0287

Joshi, A. K. (1985). "Tree adjoining grammars: How much context sensitivity is required to provide reasonable structural descriptions?," in *Natural language parsing. psychological, computational, and theoretical perspectives*, eds D. R. Dowty, L. Karttunen, and A. M. Zwicky (Cambridge: Cambridge University Press), 206–250. doi: 10.1017/CBO9780511597855.007

Kahnemuyipour, A. (2004). *The syntax of sentential stress*. Ph.D. thesis. Toronto: University of Toronto.

Kahnemuyipour, A. (2009). *The syntax of sentential stress*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199219230.001.0001

Katz, J., and Pesetsky, D. (2011). *The identity thesis for language and music*. Available online at: http://ling.auf.net/lingbuzz/000959 (accessed August 01, 2022).

Kemmerer, D. (2012). The cross-linguistic prevalence of SOV and SVO word orders reflects the sequential and hierarchical representation of action in Broca's area. *Lang. Linguist. Compass* 6, 50–66. doi: 10.1002/lnc3.322

Kemmerer, D. (2015). Word order, action, and the brain: A reply to Arbib. *Lang. Linguist. Compass* 9, 150–156. doi: 10.1111/lnc3.12132

Kemmerer, D. (2021). What modulates the mirror neuron system during action observation? Multiple factors involving the action, the actor, the observer, the relationship between actor and observer, and the context. *Prog. Neurobiol.* 205:102128. doi: 10.1016/j.pneurobio.2021.102128

Koelsch, S. (2012). *Brain and music*. Hoboken, NJ: Wiley-Blackwell.

Koelsch, S., Gunter, T. C., von Cramon, D. Y., Zysset, S., Lohmann, G., and Friederici, A. D. (2002). Bach speaks: A cortical "language-network" serves the processing of music. *Neuroimage* 17, 956–966. doi: 10.1006/nimg.2002.1154

Kotilahti, K., Nissilä, I., Näsi, T., Lipiäinen, L., Noponen, T., Meriläinen, P., et al. (2010). Hemodynamic responses to speech and music in newborn infants. *Hum. Brain Mapp.* 31, 595–603. doi: 10.1002/hbm.20890

Kratzer, A., and Selkirk, E. (2007). Phase theory and prosodic spell-out: The case of verbs. *Linguist. Rev.* 24, 93–135. doi: 10.1515/TLR.2007.005

Ladd, D. R. (2008). *Intonational phonology*, 2nd Edn. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511808814

Langus, A., Marchetto, E., Bion, R. A. H., and Nespor, M. (2012). Can prosody be used to discover hierarchical structure in continuous speech? *J. Mem. Lang.* 66, 285–306. doi: 10.1016/j.jml.2011.09.004

Lashley, K. (1951). "The problem of serial order in behavior," in *Cerebral mechanisms in behavior: The Hixon symposium*, ed. L. A. Jeffress (New York, NY: Wiley), 112–147.

Leonardi, S., Cacciola, A., De Luca, R., Aragona, B., Andronaco, V., Milardi, D., et al. (2017). The role of music therapy in rehabilitation: Improving aphasia and beyond. *Int. J. Neurosci.* 128, 90–99. doi: 10.1080/00207454.2017.1353981

Liddell, S. K. (1978). "Nonmanual signals and relative clauses in American sign language," in *Understanding language through sign language research*, ed. P. Siple (New York,NY: Academic Press).

Liddell, S. K. (1980). *American sign language syntax*. The Hague: Mouton. doi: 10.1515/9783112418260

Lieberman, P. (2006). *Toward an evolutionary biology of language*. Cambridge, MA: Harvard University Press. doi: 10.4159/9780674274839

Liu, F., Patel, A. D., Fourcin, A., and Stewart, L. (2010). Intonation processing in congenital amusia: Discrimination, identification and imitation. *Brain* 133, 1682–1693. doi: 10.1093/brain/awq089

Liu, W. C., Gardner, T. J., and Nottebohn, F. (2004). Juvenile zebra finches can use multiple strategies to learn the song. *Proc. Natl. Acad. Sci. U.S.A.* 101, 18177–18182. doi: 10.1073/pnas.0408065101

MacLarnon, A. M., and Hewitt, G. P. (1999). The evolution of human speech: The role of enhanced breathing control. *Am. J. Phys. Anthropol.* 109, 341–363. doi: 10.1002/(SICI)1096-8644(199907)109:3<341::AID-AJPA5<3.0.CO;2-2

Marin, O. (1989). Neuropsychology, mental cognitive models, and music processing. *Contemp. Music Rev.* 4, 255–263. doi: 10.1080/07494468900640341

Marler, P. (1998). "Animal communication and human language," in *The origin and diversification of language. Wattis symposium series in anthropology. Memoirs of the California academy of sciences, No. 24*, eds G. Jablonski and L. C. Aiello (San Francisco, CA: California Academy of Sciences), 1–19.

Marler, P. (2000). "Origins of music and speech: Insights from animals," in *The origins of music*, eds N. Wallin, B. Merker, and S. Brown (London: The MIT Press), 31–48.

Martinez, I., Rosa, M., Quinn, R., Jarabo, P., Lorenzo, C., Bonmati, A., et al. (2013). Communicative capacities in Middle Pleistocene, humans from the Sierra de Atapuerca in Spain. *Quaternary Int.* 295, 94–101. doi: 10.1016/j.quaint.2012.07.001

Meyer, M., Alter, K., Friederici, A. D., Lohmann, G., and von Cramon, D. Y. (2002). fMRI reveals brain regions mediating slow prosodic modulations in spoken sentences. *Human Brain Mapp.* 17, 73–88. doi: 10.1002/hbm.10042

Miller, G. A., Galanter, E., and Pribram, K. H. (1960). *Plans and the structure of behavior*. New York, NY: Holt, Rinehart and Winston, Inc. doi: 10.1037/10039-000

Mithen, S. (2005). *The singing neanderthals: The origins of music, language, mind and body*. London: Weidenfeld and Nicolson.

Miyagawa, S. (2017). "Integration hypothesis: A parallel model of language development in evolution," in *Evolution of the brain, cognition, and emotion in vertebrates*, eds S. Watanabe, M. Hofman, and T. Shimizu (New York, NY: Springer), 225–247. doi: 10.1098/rstb.2013.0298

Miyagawa, S., Berwick, R. C., and Okanoya, K. (2013). The emergence of hierarchical structure in human language. *Front. Psychol.* 4:71. doi: 10.3389/fpsyg.2013.00071

Miyagawa, S., Ojima, S., Berwick, R. C., and Okanoya, K. (2014). The integration hypothesis of human language evolution and the nature of contemporary languages. *Front. Psychol.* 5:564. doi: 10.3389/fpsyg.2014.00564

Mol, C., Chen, A., Kager, R. W. J., and Haar, S. M. (2017). Prosody in birdsong: A review and perspective. *Neurosci. Biobehav. Rev.* 81, 167–180. doi: 10.1016/j.neubiorev.2017.02.016

Moore, M. W. (2010). "Grammars of action' and stone flaking design space," in *Stone tool and the evolution of human cognition*, eds A. Nowell and I. Davidson (Boulder, CO: University Press of Colorado), 13–43.

Moorman, S., and Bolhuis, J. J. (2013). "Behavioral similarities between birdsong and spoken language," in *Birdsong, speech, and language: Exploring the evolution of mind and brain*, eds J. J. Bolhuis and M. Everaert (Cambridge, MA: MIT Press), 111–123. doi: 10.7551/mitpress/9322.003.0009

Morley, I. (2013). *The prehistory of music: Human evolution, archeology, and the origins of musicality*. Oxford: Oxford University Press. doi: 10.1080/00293652.2014.949838

Nan, Y., Sun, Y. N., and Peretz, I. (2010). Congenital amusia in speakers of a tone language: Association with lexical tone agnosia. *Brain* 133, 2635–2642. doi: 10.1093/brain/awq178

Neidle, C., Kegl, J., Maclaughlin, D., Bahan, B., and Lee, R. (2000). *The syntax of American sign language: Functional categories and hierarchical structure*. Cambridge, MA: MIT Press.

Nevue, A. A., Lovell, P. V., Wirthlin, M., and Mello, C. V. (2020). Molecular specializations of deep cortical layer analogs in songbirds. *Sci. Rep.* 10:18767. doi: 10.1038/s41598-020-75773-4

Nooteboom, S. (1997). The prosody of speech: Melody and rhythm. *Handb. Phon. Sci.* 5, 640–673.

O'Grady, W. (2005). *How children learn language*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511791192

Oesch, N. (2019). Music and language in social interaction: Synchrony, antiphony and functional origins. *Front. Psychol.* 10:1514. doi: 10.3389/fpsyg.2019.01514

Oesch, N. (2020). "Evolutionary musicology," in *Encyclopedia of evolutionary psychological science*, eds T. A. Shackelford and V. Weeks-Shakelford (London: Springer), 2725–2729. doi: 10.1007/978-3-319-16999-6_2845-1

Osiurak, F., Lasserre, S., Arbanti, J., Brogniart, J., Bluet, A., Navarro, J., et al. (2021). Technical reasoning is important for cumulative technological culture. *Nat. Hum. Behav.* 5, 1643–1651. doi: 10.1038/s41562-021-01159-9

Palomero-Gallagher, N., and Zilles, K. (2019). Differences in cytoarchitecture of Broca's region between human, ape, and macaque brains. *Cortex* 118, 132–153. doi: 10.1016/j.cortex.2018.09.008

Patel, A. D. (2008). *Music, language, and the brain*. New York, NY: Oxford University Press.

Patel, A. D. (2010). "Music, biological evolution, and the brain," in *Emerging disciplines: Shaping new fields of scholarly inquiry in and beyond the humanities*, ed. M. Bailar (Houston, TX: OpenStax CNX), 41–64.

Patel, A. D. (2012). "Language, music, and the brain: A resource-sharing framework," in *Language and music as cognitive systems*, eds P. Rebuschat, M. Rohrmeier, J. A. Hawkins, and I. Cross (Oxford: Oxford University Press), 204–223. doi: 10.1093/acprof:oso/9780195123753.001.0001

Patel, A. D. (2021). Vocal learning as a preadaptation for the evolution of human beat perception and synchronization. *Philos. Trans. R. Soc. B Biol. Sci.* 376:20200326. doi: 10.1098/rstb.2020.0326

Patel, A. D., Gibson, E., Ratner, J., Besson, M., and Holcomb, P. J. (1998). Processing syntactic relations in language and music: An event-related potential study. *J. Cogn. Neurosci.* 10, 717–733. doi: 10.1162/089892998563121

Peretz, I. (1993). Auditory atonalia for melodies. *Cogn. Neuropsychol.* 10, 21–56. doi: 10.1080/02643299308253455

Peretz, I., and Morais, J. (1989). Music and modularity. *Contemp. Music Rev.* 4, 277–291. doi: 10.1080/07494468900640361

Peretz, I., and Morais, J. (1993). "Specificity for music," in *Handbook of neuropsychology*, eds F. Boller and J. Grafman (Amsterdam: Elsevier), 373–390.

Peretz, I., Kolinsky, R., Tramo, M., Labrecque, R., Hublet, C., Demeurisse, G., et al. (1994). Functional dissociations following bilateral lesions of auditory cortex. *Brain* 117, 1283–1301. doi: 10.1093/brain/117.6.1283

Pfenning, A. R., Hara, E., Whitney, O., Rivas, M. V., Wang, R., Roulhac, P. L., et al. (2014). Convergent transcriptional specializations in the brain of humans and song-learning birds. *Science* 346:1256846. doi: 10.1126/science.1256846

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., and Fong, C. (1991). The use of prosody in syntactic disambiguation. *J. Acoust. Soc. Am.* 90, 2956–2970. doi: 10.1121/1.401770

Putt, S. S., Wijeakumar, S., Franciscus, R. G., and Spencer, J. P. (2017). The functional brain networks that underlie Early Stone Age tool manufacture. *Nat. Hum. Behav.* 1:0102. doi: 10.1038/s41562-017-0102

Rauschecker, J. P., and Scott, S. (2009). Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nat. Neurosci.* 12, 718–724. doi: 10.1038/nn.2331

Reinhart, T. (2006). *Interface strategies: Optimal and costly computations*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/3846.001.0001

Richards, N. (2010). *Uttering trees*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9780262013765.001.0001

Richards, N. (2016). *Contiguity theory*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/10542.001.0001

Richards, N. (2017). *Deriving contiguity*. Cambridge, MA: MIT.

Roberts, I. (2012). "Comments and a conjecture inspired by Fabb and Halle," in *Language and music as cognitive systems*, eds P. Rebuschat, M. Rohrmeier, J. A. Hawkins, and I. Cross (Oxford: Oxford University Press), 51–66. doi: 10.1093/acprof:oso/9780199553426.003.0003

Rogalsky, C., Rong, F., Saberi, K., and Hickok, G. (2011). Functional anatomy of language and music perception: Temporal and structural factors investigated using functional magnetic resonance imaging. *J. Neurosci.* 31, 3843–3852. doi: 10.1523/JNEUROSCI.4515-10.2011

Sammler, D., Kotz, S. A., Eckstein, K., Ott, D. V., and Friederici, A. D. (2010). Prosody meets syntax: The role of the corpus callosum. *Brain* 133, 2643–2655. doi: 10.1093/brain/awq231

Samuels, B. D. (2015). Can a bird brain do phonology? *Front. Psychol.* 6:1082. doi: 10.3389/fpsyg.2015.01082

Sandler, W., Meir, I., Dachkovsky, S., Padden, C., and Aronoff, M. (2011). The emergence of complexity in prosody and syntax. *Lingua* 121, 2014–2033. doi: 10.1016/j.lingua.2011.05.007

Schafer, A. J., Speer, S. R., Warren, P., and White, S. D. (2000). Intonational disambiguation in sentence production and comprehension. *J. Psycholinguist. Res.* 29, 169–182. doi: 10.1023/A:1005192911512

Schenker, N. M., Hopkins, W. D., Spocter, M. A., Garrison, A. R., Stimpson, C. D., Erwin, J. M., et al. (2010). Broca's area homologue in chimpanzees (*Pan troglodytes*): Probabilistic mapping, asymmetry, and comparison to humans. *Cereb. Cortex* 20, 730–742. doi: 10.1093/cercor/bhp138

Schön, D., Gordon, R., Campagne, A., Magne, C., Astésano, C., Anton, J. L., et al. (2010). Similar cerebral networks in language, music and song perception. *Neuroimage* 51, 450–461. doi: 10.1016/j.neuroimage.2010.02.023

Selkirk, E. (1974). French Liaison and the $\bar{X}$-notation. *Linguist. Inq.* 5, 573–590.

Selkirk, E. (1986). On derived domains in sentence phonology. *Phonology Yearbook* 3, 371–405. doi: 10.1017/S0952675700000695

Selkirk, E. (1995). "Sentence prosody: Intonation, stress, and phrasing," in *The handbook of phonological theory*, ed. J. A. Goldsmith (London: Blackwell), 550–569.

Selkirk, E. (2009). On clause and intonational phrase in Japanese: The syntactic grounding of prosodic constituent structure. *Gengo Kenkyu* 136, 35–73.

Selkirk, E. (2011). "The syntax–phonology interface," in *The handbook of phonological theory*, 2nd Edn, eds J. A. Goldsmith, J. J. Riggle, and A. C. L. Yu (Oxford: Wiley-Blackwell), 435–484. doi: 10.1002/9781444343069.ch14

Sergent, J. (1993). Mapping the musician brain. *Hum. Brain Mapp.* 1, 20–38. doi: 10.1002/hbm.460010104

Shilton, D. (2022). Sweet participation: The evolution of music as an interactive technology. *Music Sci.* 5, 1–15. doi: 10.1177/20592043221084710

Smaers, J. B., Gómez-Robles, A., Parks, A. N., and Sherwood, C. C. (2017). Exceptional evolutionary expansion of prefrontal cortex in great apes and humans. *Curr. Biol.* 27, 714–720. doi: 10.1016/j.cub.2017.01.020

Sparks, R., Helm, N., and Albert, M. (1974). Aphasia rehabilitation resulting from melodic intonation therapy. *Cortex* 10, 303–316. doi: 10.1016/S0010-9452(74)80024-9

Speer, S. R., Warren, P., and Schafer, A. J. (2011). Situationally independent prosodic phrasing. *Lab. Phonol.* 2, 35–98. doi: 10.1515/labphon.2011.002

Stout, D. (2011). Stone toolmaking and the evolution of human culture and cognition. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 366, 1050–1059. doi: 10.1098/rstb.2010.0369

Stout, D., and Chaminade, T. (2012). Stone tools, language and the brain in human evolution. *Philos. Trans. R. Soc. B Biol. Sci.* 367, 75–87. doi: 10.1098/rstb.2011.0099

Stout, D., and Hecht, E. (2014). "Neuroarchaeology," in *Human paleoneurology springer series in bio-/neuroinformatics*, ed. E. Bruner (Cham: Springer), 145–175. doi: 10.1007/978-3-319-08500-5_7

Stout, D., Chaminade, T., Thomik, A., Apel, J., and Faisal, A. (2018). Grammars of action in human behavior and evolution. *biorXiv* 3:281543. doi: 10.1101/281543

Stout, D., Toth, N., Schick, K., and Chaminade, T. (2008). Neural correlates of early stone age toolmaking: Technology, language and cognition in human evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 363, 1939–1949. doi: 10.1098/rstb.2008.0001

Tattersall, I. (2008). An evolutionary framework for the acquisition of symbolic cognition by *Homo sapiens*. *Comp. Cogn. Behav. Rev.* 3, 99–114. doi: 10.3819/ccbr.2008.30006

Tattersall, I. (2010). Human evolution and cognition. *Theory Biosci.* 129, 193–201. doi: 10.1007/s12064-010-0093-9

Tattersall, I. (2012). *Masters of the planet: The search for our human origins*. New York, NY: Palgrave Macmillan.

Tattersall, I. (2016). Language origins: An evolutionary framework. *Topoi* 37, 1–8. doi: 10.1007/s11245-016-9368-1

Tillmann, B., Burnham, D., Nguyen, S., Grimault, N., Gosselin, N., and Peretz, I. (2011). Congenital amusia (or tone-deafness) interferes with pitch processing in tone languages. *Front. Psychol.* 2:120. doi: 10.3389/fpsyg.2011.00120

Tillmann, B., Janata, P., and Bharucha, J. J. (2003). Activation of the inferior frontal cortex in musical priming. *Brain Res. Cogn. Brain Res.* 16, 145–161. doi: 10.1016/S0926-6410(02)00245-8

Truckenbrodt, H. (2006). "Phrasal stress," in *The encyclopedia of languages and linguistics*, ed. K. Brown (Oxford: Elsevier), 572–579. doi: 10.1016/B0-08-044854-2/04447-3

Weintraub, S., Mesulam, M. M., and Kramer, L. (1981). Disturbances in prosody: A right-hemisphere contribution to language. *Arch. Neurol.* 38, 742–745. doi: 10.1001/archneur.1981.00510120042004

Williams, H., and Staples, K. (1992). Syllable chunking in zebra finch (*Taeniopygia guttata*) song. *J. Comp. Psychol.* 106, 278–286. doi: 10.1037/0735-7036.106.3.278

Wynn, T. (1991). Tools, grammar and the archaeology of cognition. *Camb. Archaeol. J.* 1, 191–206. doi: 10.1017/S0959774300000354

Yip, M. J. (2006). The search for phonology in other species. *Trends Cogn. Sci.* 10, 442–446. doi: 10.1016/j.tics.2006.08.001

Yip, M. J. (2013). "Structure in human phonology and in birdsong: A phonologist's perspective," in *Birdsong, speech, and language: Exploring the evolution of mind and brain*, eds J. J. Bolhuis and M. Everaert (Cambridge, MA: MIT Press), 181–208. doi: 10.7551/mitpress/9322.001.0001

Zaccarella, E., and Friederici, A. D. (2015a). Reflections of word processing in the insular cortex: A sub-regional parcellation based functional assessment. *Brain Lang.* 142, 1–7. doi: 10.1016/j.bandl.2014.12.006

Zaccarella, E., and Friederici, A. D. (2015b). Merge in the human brain: A sub-region based functional investigation in the left pars opercularis. *Front. Psychol.* 6:1818. doi: 10.3389/fpsyg.2015.01818

Zaccarella, E., and Friederici, A. D. (2015c). "Syntax in the brain," in *Brain mapping: An encyclopedic reference*, ed. A. W. Toga (Cambridge, MA: Academic Press), 461–468. doi: 10.1016/B978-0-12-397025-1.00268-2

Zubizarreta, M. L. (1998). *Prosody, focus and word order*. Cambridge, MA: MIT Press.

Zubizarreta, M. L. (2009). The syntax and prosody of focus: The Bantu-Italian connection. *Iberia* 2, 1–39. doi: 10.3389/fpsyg.2018.00059