universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

"The musical origins of language: prosodic protolanguage and the link to modern communication"

verfasst von / submitted by

Christos Zervas

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien, 2021 / Vienna, 2021

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066812
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium English Language and Linguistics
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Nikolaus Ritt

Table of contents

List of figures ..... *i*

List of abbreviations ..... *i*

Acknowledgements ..... *ii*

English Abstract ..... *ivv*

German Abstract ..... *v*

1. Introduction ***1***

2. The plausibility of a common precursor of language and music ***4***

2.1. Shared properties and differences between language and music 4

2.1.1 Pitch, timbre, and melody 5

2.1.2. Rhythm 7

2.1.3. Syntax 9

2.1.4. Meaning 11

2.2. Cognitive and neural processing 12

2.2.1. Brain imaging studies 12

2.2.2. Processing of prosodic patterns 15

2.2.3. Experimental evidence 18

2.3. Developmental evidence 20

2.3.1. Innate predispositions to language and music 20

2.3.2. Prosodic scaffolding 23

2.4. Cross-species comparative data 26

2.4.1. Nonhuman primates 27

2.4.2. Birds and non-primate mammals 30

3. Musical protolanguage proposals ***33***

3.1. Selective pressures 34

3.1.1. Sexual selection 34

3.1.2. Kin Selection 41

3.1.3. Social selection 48

3.2. The “musilanguage” model 57

3.3. Holistic protolanguage 59

4. Objections to musical protolanguage ***..... 64***

5. The emergence of language: reconstruction of its evolutionary history..... 69

5.1. Gradualism 69

5.2. Archeological record 70

5.2.1. Paleontological data 71

5.2.2. Paleogenetic data 74

5.3. Evolutionary steps in the emergence of language 76

6. Conclusion 79

7. References 81

List of figures

Figure 1. Wernicke and Broca's areas.....13
Figure 2. Evolutionary relationships between extant primates.....29
Figure 3. Hominin species mentioned in the text and their approximate time spans...44
Figure 4. Location of the hypoglossal canal in the base of the human skull.....72
Figure 5. Thoracic vertebral canal.....73

List of abbreviations

ADS	adult-directed speech
BA	Brodmann's area
fMRI	functional magnetic resonance imaging
Hz	hertz
IDS	infant-directed speech
L1	first language
LCA	last common ancestor
MEG	magnetoencephalography
MRI	magnetic resonance imaging
ms	millisecond
MYA	million years ago
NP	noun phrase
OED	Oxford English Dictionary
PET	positron emission tomography
VP	verb phrase

Acknowledgements

Dedicated to my wife Xuejiao. Her infinite encouragement fills me with strength and optimism.

I am also grateful to my supervisor Nikolaus Ritt for his support and guidance.

English Abstract

Language and music define us as humans and gaining insights on how they emerged is essential to understanding ourselves. A close examination of the relationship between language and music shows that the two domains have more in common than typically thought. An increasing capacity for vocal control and vocal learning were decisive in the evolution of language and musicality. Although language has been traditionally assumed to be a monolithic entity, current evidence demonstrates that it is multi-faceted and evolved gradually over a long period. Findings from a broad range of disciplines point to the importance of prosody in communication and its ancient provenance in the evolution of language. The ability to volitionally regulate pitch contour, timbre, amplitude, and duration in vocalizations was crucial in the first steps towards the emergence of language. Thus, a prosodic/musical protolanguage anticipated the appearance of syntax, which was a more recent addition in human communicative systems. A multicomponent approach to language evolution helps us understand its complex nature and acknowledge the interdisciplinary character of the problem. Empirical and experimental evidence from linguistics, musicology, psychology, primatology, paleontology, cognitive science, and genetics converge on the notion of a common precursor of language and music which functioned as an intermediate system between nonhuman primate vocalizations and modern analytic language. The timescale that was required for language to evolve suggests that its evolutionary trajectory has been influenced by multiple selective pressures. Given that humans and most other primates are deeply social creatures, complex social structure must have been a major selective force. The musical protolanguage hypothesis challenges traditional ways of thinking about the nature of language and brings prosody to the foreground of communicating not only affective but also linguistic meaning.

German Abstract (Deutsche Zusammenfassung)

Sprache und Musik bestimmen uns als Menschen und es ist erforderlich, dass wir Einblick gewinnen, in wie sie entstanden sind, um uns selbst verstehen zu können. Eine eingehende Überprüfung des Verhältnisses zwischen Sprache und Musik weist darauf hin, dass die zwei Bereiche mehr gemeinsam haben als ursprünglich gedacht. Die ansteigende Fähigkeit zu Stimmkontrolle und Erlernen der Stimme waren entscheidend für die Entwicklung von Sprache und Musikalität. Obwohl man traditionell davon ausging, dass Sprache ein monolithisches Gebilde ist, zeigen aktuelle Erkenntnisse, dass sie vielschichtig ist und sich über einen langen Zeitraum hinweg allmählich entwickelt hat. Der Einblick aus einer breiten Auswahl an Disziplinen zeigt die Bedeutung der Prosodie in der Kommunikation und ihre uralte Herkunft aus der Evolution der Sprache an. Die Fähigkeit, Tonhöhenkontur, Klangfarbe, Amplitude und Dauer von Vokalisierungen willensmäßig zu regeln, war ausschlaggebend für die ersten Schritte zur Entstehung von Sprache. Daher nahm eine prosodisch/musikalische Protosprache das Auftreten des Satzbaus an, der kürzlich in die menschlichen Kommunikationssysteme aufgenommen wurde. Ein Mehrkomponenten-Ansatz für die Sprachevolution hilft uns dabei, ihre komplexe Natur zu verstehen und den interdisziplinären Charakter des Problems anzuerkennen. Empirische und experimentelle Hinweise aus den Bereichen Linguistik, Musikwissenschaft, Psychologie, Primatologie, Paläontologie, Kognitionswissenschaft und Genetik deuten auf die Idee eines gemeinsamen Vorläufers von Sprache und Musik hin, der als Zwischensystem zwischen den Vokalisationen von nichtmenschlichen Primaten und der modernen analytischen Sprache diente. Der für die Entwicklung der Sprache nötige Zeitraum deutet an, dass ihr Entwicklungsverlauf von mehreren Selektionsdrücken beeinflusst wurde. Da der Mensch und die meisten anderen Primaten zutiefst soziale Lebewesen sind, muss die komplexe Sozialstruktur ein wichtiger Selektionsfaktor gewesen sein. Die Annahme der musikalischen Protosprache stellt traditionelle Denkweisen über die Natur der Sprache in Frage und bringt die Prosodie in den Vordergrund der Kommunikation von nicht nur affektiver, sondern auch sprachlicher Bedeutung.

1. Introduction

Human language and music are two defining traits of our species, yet no one can tell with certainty when, how, or why they emerged, and what evolutionary path they have followed. The questions of the origins of language and music are fundamental to understanding our species and they have attracted the interest of researchers from various disciplines. Although earlier attempts to address these questions were highly speculative in nature, substantial cross-disciplinary advances in the recent years make it possible to investigate the issue scientifically by formulating testable hypotheses based on empirical grounds (see Boeckx 2017; Fitch 2017; Ravignani, Thompson & Filippi 2018).

The main concern of this thesis is the role of intonation, stress, and rhythm in human communication, and their evolutionary relevance in the emergence of language. In particular, it is argued that a prosodic/musical[1] protolanguage evolved from primate vocalizations, serving as the basis for the subsequent emergence of modern language and music. Although the importance of facial expressions, gestures, and bodily movements in communication is not denied, the role of prosody is deemed primary (cf. Corballis 2015). Furthermore, while it is acknowledged that there cannot have been a single cause that spurred the emergence of language and music, their social dimension is emphasized: the two domains could not have emerged independently of the social context in which they occur.

The putative protolanguage was dependent on prosodic properties of utterances (intonation, rhythm, degree of loudness, stress) rather than on syntax to achieve communicative ends. Thus, the thesis advanced here comes in opposition with traditional models of language evolution that argue for a 'syntax first' scenario (cf. Bickerton 2000). An enhanced capacity for *voice modulation*[2] and *vocal control*[3] was a key evolutionary development; hominins acquired an improved ability to volitionally generate vocal signals regulated for pitch, contour, intensity, and duration (e.g., Morley

[1] *Prosodic protolanguage* and *musical protolanguage* are used interchangeably. The term *musical* does not imply any formal musical structures.
[2] *Voice modulation* is the manipulation of nonverbal properties of the voice, such as pitch and timbre (Pisanski et al. 2016: 306).
[3] *Vocal control* is "the capacity to control the larynx [...] or the supralaryngeal vocal tract [...] in a flexible or voluntary manner" (Pisanski et al. 2016: 305).

2013). It needs to be stressed that vocal control was intentional and voluntary. This is important, because it allowed our ancestors to produce vocal expressions that were not dependent on psychological or physiological states of arousal. Vocal learning was also crucial; the ability for "voco-mimesis", possibly stemming from a more general cognitive capacity for motor learning, was a precondition for the emergence of syntactic language (Donald 2017: 204). The proposed musical protolanguage may initially have been an iconic system of communication, later acquiring arbitrary and symbolic meanings. As such, it provided the basis for the development of language as we know it today.

There have been various evolutionary accounts regarding the origins of language and music. Several authors have argued in favor of a common precursor of the two traits, which has been variably dubbed 'musical protolanguage', 'prosodic protolanguage', or 'protomusic' (Masataka 2009: 11; Fitch 2013: 498-499; Fenk-Oczlon 2017: 1), but there is no consensus on the nature of the presumed protolanguage. For some theorists, music in the form of wordless singing anticipated the emergence of language (e.g., Dunbar 2017). The prosodic/musical protolanguage hypothesis discussed here stresses the importance of prosodic modulation of the voice in the evolution of human speech and language. It does not entail that music preceded language, or that our distant ancestors were singing creatures before they became linguistic creatures. The capacity for modern musical skills, which require vocal imitation, rhythm perception, beat synchronization, melodic contour processing, and hierarchical embedding probably evolved together with the full suite of modern linguistic capacity (Morley 2013: 225; de Boer 2017: 160; Fitch 2017: 24; Oesch 2019: 2).

I draw on findings from a range of fields that present evidence in favor of the plausibility of the prosodic protolanguage hypothesis. First, the relationship between language and music is examined (section 2.1), followed by findings on the cognitive processing of the two domains (section 2.2). Section 2.3 presents insights from developmental psychology and infant perceptual abilities. In section 2.4, data from nonhuman primate and other animal communication systems are considered. Proposals that deal with the selective pressures thought to have been relevant for the emergence of language and/or music are assessed in section 3.1, while assumptions about the nature of the putative protolanguage are discussed in sections 3.2 and 3.3. Section 4 addresses objections to the plausibility of a musical protolanguage. In section 5, a

tentative attempt is made to reconstruct the evolutionary history of language, building on the discussion that precedes it and, additionally, on evidence from the fossil record and paleogenetic data. In section 6 I bring together the consequences of all sources of evidence and I end with my conclusions.

By investigating how language originated, I focus on its biological evolution; biological evolution can only explain the ability to acquire a language, and it is different from the cultural evolution of specific languages, which explains language variability and diversification (see Fitch 2017: 12). Similarly, the evolution of the capacity for musicality should not be confused with "the practice of music as a learned cultural behaviour" (Morley 2013: 278). Nonetheless, it is acknowledged that the two processes are not mutually exclusive; cultural processes may influence genetic variation (Laland, Odling-Smee & Myles 2010), and thus a gene-culture interplay in the evolution of the language faculty cannot be excluded. In this thesis, I concentrate on the biological evolution of language, which is regarded as a complex adaptive trait that evolved in small incremental steps.[4] Hence, single mutation events or 'saltationist' scenarios are rejected on the grounds of being evolutionarily implausible (see section 5.1).

I am aware that language and music have been defined in many different ways. Language is traditionally considered a cognitive system associated with thought, whereas speech is the physical manifestation of that system (see Botha 2009: 65-66; de Boer 2005: 281). According to Oxford English Dictionary (OED), language is "[t]he system of spoken or written communication [...] typically consisting of words used within a regular grammatical and syntactic structure" (OED: s.v. *language, n*). Here, language and speech are taken to be inextricably linked, and the evolution of language is thought to be contingent on the evolution of speech (cf. Chomsky 2007). Fully-fledged *language* pertains to modern lexical language, characterized by referentiality, propositional meaning, and hierarchical structure. *Protolanguage* refers to any intermediate system between nonhuman primate vocalizations and modern language. Therefore, a *protolanguage* is more expressive than nonhuman primate vocal signals, but it lacks some of the defining elements of modern language. *Speech* is deployed as a more

[4] See also "complex adaptive system" (Ritt 2004: 92). Complex adaptive systems have an ability to evolve and alter their structure in response to specific environmental influences without being under the control of a central agent; language and human cognition may be treated as such systems (Ritt 2004: 91-92).

general, inclusive term: it may refer to the physical instantiation of any protolinguistic system besides modern language, and it also includes the tonal aspects of communication.

The term *music* has been considerably harder to define than language (see Nettl 2000: 472). The lack of absolute musical universals (Savage et al. 2015) means that definitions of music are bound to be shaped by cultural preconceptions. OED defines human music as "[t]he art or science of combining vocal or instrumental sounds to produce beauty of form, harmony, melody, rhythm, expressive content" (OED: s.v. *music, n*). Like language, human music differs from other animal vocalizations, such as the songs of birds, in that it has metrical hierarchy (see section 2.1.2) and is under our volitional control. I use the adjective *musical* in a wider sense to include rhythmical sounds or vocal modulations that may not necessarily be characterized by hierarchical structures of sound sequences or occur in a voluntary manner. Therefore, musical behaviors *sensu lato* are not exclusive to humans. It should be noted though that I do not attempt to draw a clear line between musical and non-musical behaviors, or between language and protolanguage. Evolution is characterized by gradualism and continuity, and thus I see no compelling reason in trying to identify a particular point in time in which language or music originated; there was no single moment that such a transition took place (see section 5.1).

2. The plausibility of a common precursor of language and music

2.1. Shared properties and differences between language and music

Language and music are typically regarded as two different cognitive domains. Despite the differences, there are many interesting parallels:

(a) Human language and music are two characteristics unique in our species that do not have a parallel in the animal kingdom.

(b) They are both expressed through the medium of sound, i.e., through speech and song. Although much of contemporary music is instrumental, comparative work in ethnomusicology makes clear that the human voice has a primary role

compared to instruments (Morley 2013: 30-32, 309). Hence, the vocal tract is central in the expression of both language and music.

(c) Vocal learning and elaborate vocal control are necessary in the production of both language and music (Jackendoff 2009: 196-197).

(d) Both traits are often combined, as they are in poetry, lyrical song, and chant.

(e) They are present in every human culture, i.e., they are human universals.

(f) They are culturally learned and transmitted.

Therefore, any evolutionary theory of either trait will have implications for the other (Masataka 2009: 11). Their relationship, including their formal structure, is examined in more detail below, focusing on their structural and acoustic properties.

2.1.1 Pitch, timbre, and melody

Pitch refers to the quality of a sound as it is determined by the vibration rate; in the case of animals that possess a larynx (like humans), voice pitch quality is related to the vibration rate of the vocal folds, which is termed *fundamental frequency* (Pisanski et al. 2016: 305). *Timbre* denotes the character of a sound that is distinct from its pitch, duration, or intensity; our perception of voice timbre depends on "resonant frequencies of the supralaryngeal vocal tract" known as *formants* (Pisanski et al. 2016: 305). Essentially, timbre is "the primary parameter that allows us to discriminate between different vowels" (Fenk-Oczlon 2017: 1), or to distinguish the sound of a cello from the sound of a violin. Although the term *melody* is normally reserved for music, an analog in speech would be the intonational patterns of an utterance. The "intonation, rhythm, and relative loudness and timing of components of an utterance" make up speech *prosody* (Nygaard, Herold & Namy 2009: 128; see also Levinson 2016: 7-8).

Scale intervals vary across musical systems (typically 1 to 3 semitones),[5] but even within a musical system, scales tend to have uneven intervals, i.e., they tend to be asymmetric (Patel 2008: 19-21). Hence, an unambiguous tonal center is important in most musical systems. This shows that humans have auditory predispositions for forming categories. In language this is typically illustrated by categorizing and

[5] In Western music, an *octave* is a series of eight consecutive notes where the last note has twice the frequency of the first. An octave is divided into 12 equal intervals known as *semitones* (Patel 2008: 13-15).

contrasting the sounds of a phonological system (e.g., voiced versus voiceless consonants) (Patel 2008: 22). A tonal center and a fixed space between pitches are thought to be irrelevant in languages; instead of discrete pitch changes, prosodic contours in speech "usually involve a continuous rise and fall" (Jackendoff 2009: 199). However, tone languages comprise more than half of the world's languages (Brown 2017: 11; Fenk-Oczlon 2017: 2; Patel 2008: 40). In tone languages "vowels carry the pitch modulations that convey grammatical and lexical information" (Fenk-Ocalon 2017: 2), so that "changing the pitch can completely change the meaning of the word" (Patel 2008: 40). Additionally, in most of the languages that are tonal, level tones are much more common than contour tones[6] (Patel 2008: 41): the average intertone interval in such languages is 2 to 3 semitones (Patel 2008: 42), which closely resembles the pitch spacing of musical scales across the world (Patel 2008: 19). Furthermore, melodic contours (i.e., "the patterns of ups and downs [of pitch] without regard to precise interval size") are important not only in speech prosody, but also in recognizing melodies in music (Patel 2008: 23). Thus, pitch contrasts and contours are involved in both music and language, and they are transposable: we can understand the same utterance spoken by an adult or a little child, as we can understand the same melody sung by different people or played in different instruments, because we can perceive the relations between pitches irrespective of their absolute frequencies (Fitch 2006: 179).

This does not mean that there are no differences between musical melody and speech prosody. A tonal center, pitch hierarchies, precise and stable pitch intervals, and tonality relations based on the structuring of intervals are features of musical melodies but not of speech intonation (Patel 2008: 183-185, 201, 205). The linguistic tones, whether high or low, do not have a fixed pitch level, but "relative positions within a pitch range" (Patel 2008: 209; see also Fitch 2006: 179). Despite the differences, the intonation patterns in speech may influence the tonal patterns in a culture's instrumental music; this means that the statistical learning of one domain's patterns may be transposed on another domain (Patel 2008: 218-225). Moreover, in some cultures the so-called "whistled speech" and "talking drums" are employed to exchange

[6] A *level tone* is a tone that stays at a roughly even pitch, whereas the pitch in a *contour tone* goes up or down.

messages by mimicking the linguistic tones and rhythms, thereby functioning as a "speech surrogate" (Patel 2008: 48-50; see also Fenk-Oczlon & Fenk 2009: 210). The use of such speech surrogates in certain tone languages demonstrates the significance of intonation patterns in efficient communication, even when one does not use one's own voice.

Pitch and timbre modulation occurs in both language and music, conveying a wide range of emotional information such as happiness, sadness, fear, or anger (Patel 2008: 40, 345-346; Ravignani, Thompson & Filippi 2018: 2). Even though timbral contrasts are normally associated with speech and not so much with music, one needs to consider timbre modulation in the musical traditions of small-scale indigenous cultures. For example, the Australian aboriginal wind instrument *didgeridoo* creates "a low single-tone droning sound, the 'colour' of which can be varied by changing the shape of the mouth" (Morley 2013: 24). In addition, empirical work on the association between linguistic and musical timbral contrasts indicates that there are strong links "between the musical and linguistic timbres of a culture" (Patel 2008: 67). Furthermore, cross-cultural studies demonstrate that there is a correspondence between vowel systems and musical scales (Fenk-Oczlon & Fenk 2009: 210-212; Fenk-Oczlon 2017: 3-4). For instance, Nettl (1956: 112) reports that many indigenous American cultures with three or four vowel systems have tritonic or tetratonic musical scales respectively.[7] Crucially, Fenk-Oczlon (2017: 2) points out that, in addition to timbre, vowels differ in their intrinsic pitch as well. By way of illustration, the vowel [i] differs from the vowel [o] not only in its timbre, but also in its higher intrinsic pitch. In support of this, Fenk-Oczlon and Fenk (2009: 212-213) provide empirical evidence from Alpine yodelers, showing that the melody ascends in almost all [o] → [i] successions, whereas it descends in [i] → [o] successions. Hence, pitch and timbre are more tightly interconnected than one initially might think.

2.1.2. Rhythm

Patel (2008: 96) defines rhythm as "the systematic patterning of sound in terms of timing, accent, and grouping". Both music and language seem to have a metrical hierarchy

[7] A tritonic scale has three notes per octave, while a tetratonic has four.

(Jackendoff 2009: 199). In musical meter, certain beats receive more prominence than others; correspondingly, a language's intonation patterns attach more prominence on particular syllables (Patel 2006: 100). Musical rhythm is differentiated from linguistic rhythm in that it is typically characterized by an isochronous pulse (Fitch 2006: 179; Jackendoff 2009: 199; Patel 2006: 100); this is a major difference between language and music, as speech rhythm does not have a regular beat in any language (see section 2.3.2). Nevertheless, isochronous beat is not a universal feature of music. In fact, typological research on music has revealed no absolute universals whatsoever, but only statistical universals (Savage et al. 2015: 8989-8990). Although isochrony is preponderant, there are some musical traditions that have "no time markings for individual notes" resulting in "unpulsed music" (Patel 2008: 97-98).

It has been claimed that "the rhythmic organization of speech is quite analogous to that of music" (Selkirk 1984: 9) and there is plenty of evidence that grouping is critical in both music and speech perception (Patel 2008: 106-112). *Grouping* refers to the perceptual clustering of linguistic or musical sequences (Patel 2008: 106). Rhythmical grouping of musical phrases is typically associated with a descending pitch and note lengthening (Patel 2008: 108). Rhythmical grouping is also present in language, and it is comparable to musical grouping. Prosodic boundaries in language do not necessarily coincide with word boundaries; like music, prosodic phrasing results from the interaction between pitch and the lengthening of a phoneme, (Patel 2008: 110-112). Therefore, as musical rhythm is fundamental in making sense of the music, so is linguistic rhythm in segmenting and understanding speech.

Rhythm is a key aspect of music, language, and dance, which are uniquely human characteristics. Contrary to musical rhythm, speech rhythm is more difficult to pinpoint and analyze formally. According to Kotz, Ravignani, and Fitch (2018: 896), rhythm is typically perceived as a uniform concept, which is counter-productive in the study of rhythm cognition across domains. Thus, they adopt a multi-component approach, rejecting a monolithic view of rhythm. They identify four distinct subcomponents of rhythm: (a) periodicity, (b) beat extraction of an external pulse, (c) audiomotor entrainment (i.e., motor synchronization to the beat), and (d) meter (hierarchical structure, with stressed and unstressed elements) (Kotz, Ravignani & Fitch 2018: 897). These authors argue that a multi-component view of rhythm is more fruitful when

comparing music to speech. The perception of meter is contingent on the first three subcomponents of rhythm, and it seems to be a uniquely human characteristic. That linguistic prosody and music share common ground at the metrical level points to "shared neural mechanisms for building up hierarchical structure from auditory sequences" (Kotz, Ravignani & Fitch 2018: 906). Thus, the first three, more 'basic' subcomponents of rhythm (which are not unique to humans) appear to have formed the basis which allowed for the ability for "metrical hierarchical grouping of events" (Kotz, Ravignani & Fitch 2018: 905) and the subsequent emergence of language and music. Understanding which subcomponents are involved in music, language, and nonhuman animal cognition is decisive in tracing the evolutionary history of human speech and music (Kotz, Ravignani & Fitch 2018: 906).

2.1.3. Syntax

Patel (2008: 241) defines syntax as "the principles governing the combination of discrete structural elements into sequences", a definition that is applicable to both language and music. There is cross-cultural variation in both linguistic and musical syntactic features. Nevertheless, there are some syntactic traits that are shared by all languages (e.g., verb subject, object). In contrast, no absolute universals have been observed in the syntax of musical traditions; perhaps the most typical commonality is the use of a musical scale with a restricted number of tones[8] (see Patel 2008: 242). Consequently, it is evident that grammatical categories (e.g., noun) and grammatical functions (e.g., object) in language do not have any equivalent in music (Patel 2008: 263). Another discrepancy between the two domains is that linguistic syntax comprises "long-distance dependencies"; words are connected in a more complicated manner than notes, which do not exhibit similar long-distance relationships (Patel 2008: 263-264). Long-distance dependencies are possible in language because there is "headed hierarchy" in the structure of sentences (Jackendoff 2009: 200). For example, even though (in English) the subject may be separated from the verb due to an intervening relative clause or prepositional phrase, this does not prevent us from identifying it. Moreover, Jackendoff (2009: 200) observes

[8] A *tone* is a sound determined by its pitch, duration, and timbre. A *note* is the symbol in musical notation that depicts pitch and duration of a sound.

that although conventional musical patterns bear a resemblance to prefabricated linguistic patterns (such as idioms and set phrases), there is no musical counterpart to words.

Despite the dissimilarities, language and music share a number of structural features. Even though musical elements do not form the intricate relations that linguistic elements do, they are organized hierarchically. As linguistic syntax has multiple levels of organization (morphemes → words → phrases → sentences), so does musical syntax (tones → chords[9] → chord progressions) (Patel 2008: 264). Thus, at an abstract level both domains are organized similarly: "one can have the same sentence structure with different words, and the same harmonic structure with different chords [...] such as [...] if the key[10] is changed" (Patel 2008: 265). For example, replacing a subject noun with another noun does not alter the syntactic structure of a sentence, and we can have the same harmonic structure in a melody if we move all the notes one pitch class higher.

Furthermore, another important trait that language and music share is recursion (Mithen 2005: 17; Patel 2008: 265) – placing one element inside another of the same type, which in turn can be placed inside another, and so on. Thus, in language, a relative clause can be embedded (i.e., placed) within another relative clause, which is embedded inside the main clause. Jackendoff (2009: 200-201) maintains that linguistic recursive structure has no parallel in music. Although it has been pointed out that recursion is a sui generis characteristic of humans, it has been typically associated with language but not music (e.g., Hauser, Chomsky & Fitch 2002: 1569). However, recursion is a primary trait of music, embedding "small-scale patterns of tension and relaxation [...] in larger tension-relaxation patterns of identical geometry but of a longer timescale" (Patel 2008: 265). Additionally, although languages have recursive syntax in an abstract sense, there seems to be a cognitive limit on the number of embeddings a sentence can have (see Everett 2017: 222-224). The potential for recursion in language is not greater than it is in music; thus, it might have originally evolved for reasons not directly related to language (cf. Hauser, Chomsky & Fitch 2002: 1578).

[9] A *chord* is a set of notes played simultaneously, creating harmony (Patel 2008: 248).
[10] In music theory, "[a] scale and its tonal hierarchy, plus its system of chords and chord relations, defines a '*key*' or tonal region" [my emphasis] (Patel 2008: 251).

Merker, Morley, and Zuidema (2015: 3) argue that a critical shared trait between language and music is generativity. Through the power of combinatoriality, the two domains allow to produce an "infinite pattern variety" by "a finite set of elements" (Merker, Morley, and Zuidema 2015: 3). Generativity in music is confirmed cross-culturally and constitutes one of the two natural "open-ended generative systems" (the other being linguistic generativity) (Merker, Morley, and Zuidema 2015: 3-4).

In addition, Patel (2008: 265-266) stresses that the structure in both language and music is built around the context rather than on intrinsic properties of their constituent elements. For instance, the subject is always identified in relation to its verb, and the harmonic function of a chord is perceived in comparison to other chords; an adjective is identified in relation to the noun it modifies, and the salience of a tone in a melody is understood in its relation to neighboring pitches (Patel 2008: 260, 265-266). These similarities at an abstract level indicate that the two domains have many parallels in their logical structure, as they consist of many psychological properties that are not independent from each other (Patel 2008: 266).

2.1.4. Meaning

Meaning is not dissociated from syntax; rearranging the order of linguistic or musical elements can alter the intended meaning (Patel 2008: 259). Music of course differs from language in that it is not possible to translate (Levi-Strauss 1964: 18). Additionally, all modern human languages are referential and convey propositional meaning (Fitch 2006: 176-177). Languages allow their speakers to make statements, give orders, ask questions, or make promises, what Searle (2016: 32-34) describes as speech acts. Patel (2008: 301) emphasizes the ubiquity of propositional meaning in languages and adds that they enable us to talk at the meta-level (e.g., use language to talk about language). Nonetheless, music also has meaning, albeit in a different way. As Patel (2008: 302-303) notes, music is "untranslatable", yet it can be understood regardless of one's cultural background (although culture may influence musical taste). Meaning in music can be understood in a broad sense of the term, which includes association of events and the expression of emotions (see Patel 2008: 304). Musical pieces often trigger particular thoughts or memories on us, and they convey feelings of happiness, excitement, or melancholy. Voice quality, intonation, and tempo can function as cues for such emotions

not only in music, but also in speech (Patel 2008: 312, 345). In other words, the acoustic hints for emotional meaning seem to be the same for music and language (see Juslin & Laukka 2003).

In sum, the complexity of the linguistic system gives its users the capability to convey semantic meaning that is unmatched by music. Nonetheless, both systems are meaningful in a broad sense: the inherent potential for **affective** meaning in both music and speech prosody is an important link that may be telling regarding the evolutionary history of meaning (see Brown 2017). It suggests that the communication of emotional meaning may have been primary in early protolanguage stages, before the emergence of symbols paved the way for linguistic complexity. Similarities in rhythmic and syntactic organization also point to a direction of shared history between language and music, as do auditory perceptual predispositions related to pitch and timbre. Ostensibly "domain specific" aspects of language and music should be recognized as the products of a long evolutionary process whose developmental processes might have been "domain general" (see Patel 2008: 72). Similarities and differences between the two domains should not be taken at face value. Parallels and divergences should be scrutinized to assess the evolutionary relevance of the various subcomponents of language and music. What I hope to have made clear so far is that none of the properties discussed is either language- or music-specific.

2.2. Cognitive and neural processing

2.2.1. Brain imaging studies

Language and music have traditionally been thought to be localized in the brain. Language has been conventionally linked to the left hemisphere of the brain, and music to the right (Morley 2013: 181). This oversimplification might have to do with the fact that the left cerebral hemisphere seems more dominant for analytic processing (associated with language), whereas the right is deemed primary for appositional processing (associated with emotional expression and, by extension, music) (see Brust 2003: 181). Whilst it is certainly true that certain aspects of language and music are processed in specific brain areas, language and music are multifaceted and complex, and their perception and production involve neurological structures from both hemispheres

(Falk 2000: 199; Morley 2013: 181). It has been shown in section 2.1 that pitch, timbre, rhythm, and syntax are not unique properties of either language or music, but they play a role in both domains. These properties are processed in various parts of the brain, but the perception and production of music and language requires their simultaneous deployment. That is, language and music are multi-modal activities that cannot be "sharply localized and isolated in individual areas of the brain" (Morley 2013: 181). Inevitably, language and music share numerous neurological substrates (Falk 2000: 197-198; Morley 2013: 184).

Sergent et al. (1992) conducted a study on ten pianists using magnetic resonance imaging (MRI) and positron emission tomography (PET)[11] to examine their brain activity as they read, listened, and performed musical scores on keyboard. During these tasks brain areas from both hemispheres were activated (Sergent et al. 1992: 107), some of which are thought to be closely related to speech, such as Wernicke's area (Fig. 1) in the left hemisphere (cf. Falk 2000: 198; Mithen 2005: 32). Thus, Sergent et al.'s study indicates that music perception and production comprise brain areas that are adjacent to or overlap with language processing.

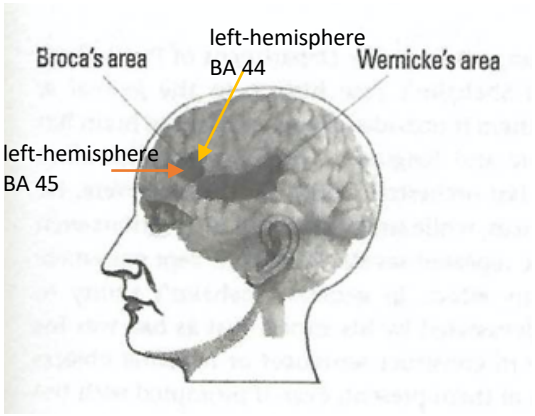

Figure 1. Wernicke and Broca's areas are located on the left cerebral hemisphere. These brain areas have been traditionally linked to linguistic abilities, as damage in them causes speech disorders (Adapted form Mithen 2005: 33).

Brown, Martinez, and Parsons (2006a) carried out a PET study to measure brain activity during music and language production. They asked amateur musicians to generate vocally melodic or linguistic phrases in response to unfamiliar musical or linguistic stimuli. Although there were lateralization tendencies (language production tended to be left-lateralized), the authors found significant overlaps. For instance,

[11] *PET* is a functional imaging technique that helps to visualize and measure metabolic activity and changes taking place in the body. It is used to evaluate the function of organs, such as the brain (John Hopkins Medicine n.d.).

sentence generation activated Brodmann's areas (BA) 44 and 45 of the left hemisphere (Brown, Martinez & Parsons 2006a: 2795). Left-hemisphere BA 44 and 45 form Broca's area (Fig. 1), which has long been identified as critical for speech production (Mithen 2005: 31-32) and is activated during syntactic processing (Brown, Martinez & Parsons 2006a: 2800; Maess et al. 2001: 543; Friederici 2017: 43). The generation of melodies also activated left-hemisphere BA 45; additionally, it activated BA 44 in the right hemisphere. Right BA 44 has also been found to be active during dancing (Brown, Martinez & Parsons 2006b: 1161). Thus, brain activity during melody generation seems to partially overlap with both speech production and dance movements.

Maess et al. (2001) used magnetoencephalography (MEG)[12] to detect the neural resources that process musical syntax. Their participants listened to harmonically related (in-key) chords and unrelated (out-of-key) chords. The unexpected, out-of-key chords violated harmonic expectancy and, as a result, caused changes in brain activity. Maess et al. (2001: 542) observed that these changes were located in area BA 44 in both hemispheres. In the left hemisphere BA 44 is part of Broca's area. Hence, Broca's area and its homologous region in the right hemisphere seem to be involved in the processing of musical syntax, at least when incongruities occur. Thus, rule-based information and sequential ordering in music and language may be processed in a similar manner. Maess et al.'s (2001: 543) study suggests that there is a tight neurological link between language and music, and that the processing of syntactic information is less domain-specific than previously thought.

Furthermore, prosody and melodic contour are two functions that point to shared substrates between language and music (Morley 2013: 187), and these functions are primarily right-hemisphere lateralized. Peretz (2003: 200) notes that right-hemisphere cerebral regions are responsible for pitch contour processing and admits that "intonation patterns of speech seem to recruit similarly located, if not identical, brain circuitries". Zatorre (2003: 240) concurs with Peretz, stressing that there seems to exist hemispheric specialization in tonal processing. Despite the seeming right-hemisphere dominance, Zatorre (2003: 237) makes clear that such putative

[12] *MEG* is a neuroimaging technique that records the magnetic fields generated by the electrical activity occurring in the brain (Singh 2014).

specializations should be treated in relative and not in absolute terms. In any case, since tonality is implicated in both music and language (see section 2.1.1.), any processing prominence in the right hemisphere is not evidence for music specialization. Furthermore, the cognitive and neural underpinnings of rhythm appear to be distributed across both cerebral hemispheres. Grahn and Brett (2007) conducted an experiment using functional magnetic resonance imaging (fMRI)[13] in which subjects (a) listened and reproduced rhythmic patterns, and (b) indicated whether two rhythms were the same or different. Grahn and Brett (2007: 902) report that during both rhythm perception and production a bilateral network of brain areas was activated.

The findings from these brain imaging studies illustrate that the processing and production of linguistic and musical properties stimulate cognitive and neural structures in various parts of the brain and in both cerebral hemispheres. Although there are lateralization tendencies, it needs to be stressed that brain structures are neurologically intertwined and not isolated entities. The processing of the various subcomponents of language and music stimulates areas from different parts of the brain, suggesting that there are no individual structures dedicated exclusively to language or music (see also Morley 2013: 196). Certain brain areas seem to be activated in both language and music processing (see Brown, Martinez, and Parsons 2006a: 2798), whereas the functional combinations of different activated areas might be domain-specific either for language or music (cf. Morley 2013: 196).

2.2.2. Processing of prosodic patterns

It has been claimed that intonation contours in language are processed differently than melodic contours in music (Peretz & Coltheart 2003). Prosody is not organized in stable pitch intervals, which is a typical feature of music. Stable pitch interval structures allow for the creation of a tonal center and hierarchical pitch relations, elements that are absent from language (Patel 2008: 183). Such seemingly unique properties of music may lead one to the conclusion that melodic contour processing is domain-specific for music (Peretz & Coltheart 2003).

[13] *MRI* is an imaging technique that creates pictures in a set moment in time, measuring only the structure of the brain. On the other hand, *fMRI* measures responses during brain activity (Farnsworth 2019).

However, empirical evidence from musically tone-deaf individuals (or amusics)[14] suggests that speech intonation and musical perception have shared processing mechanisms (Liu et al. 2010; Patel 2008: 228-238). Liu et al. (2010) tested amusics and matched controls on statement-question discrimination, and statement-question identification and imitation. Liu et al. (2010: 1686) report that "the amusic group performed significantly worse than the control group" on these tasks. The questions differed from the statements in the direction of pitch glides at the sentence-final position, without having any additional hints to meaning (Liu et al. 2010: 1689). Thus, the impairment seems to be discriminating pitch direction contrasts on the same word. Liu et al.'s study shows that pitch deficits are relevant not only for music, but also for speech processing. The average pitch discrimination threshold for amusics is around two semitones (Foxton et al. 2004). Amusic individuals do not always report speech perception deficits because there are often syntactic cues to meaning and, more importantly, "most linguistically relevant pitch movements are likely to be in excess of the pitch-direction thresholds" of amusics, whereas musical melodies typically have pitch contrasts of two semitones or less. Hence, congenital amusia is not a disorder specific to music; pitch processing may involve mechanisms used across domains. That music and speech perception are behaviorally dissociated does not mean that they are not neurally related (Patel 2008: 238).

Rhythm is typically regarded as a property of music, but as noted in section 2.1.2., speech rhythm plays an important role in effective communication. Although linguistic rhythm is largely dependent on cultural background and the phonological rules of a language (Patel 2008: 177), there is evidence for cross-cultural universals in turn-taking during conversations (Stivers et al. 2009). Speakers of all languages avoid overlapping talk and minimize silence between conversational turns; to achieve this they do not wait for their interlocutor to pause, but "they use grammar, prosody, and pragmatics to project when they can start a next turn" (Stivers et al. 2009: 10587). For example, speakers across languages use "a boosted initial pitch in questions" apart from the language-specific grammatical cues (Levinson 2016: 9). When overlaps occur, they are

[14] *Amusics* are those who are "unable either to sing in tune or detect an out-of-tune note in a melody", and this "neuro-developmental disorder of musical perception" is referred to as "congenital amusia" (Liu et al. 2010: 1682).

on average less than 100 ms long, whereas gaps are around 200 ms (Levinson 2016: 6). Although there are variations across languages, they are quantitative only, and the same explanations apply cross-linguistically. For instance, confirmation responses are delivered faster than disconfirmation responses, and information requests are slower than other questions (Stivers et al. 2009: 10587-10589). Additionally, "linguistic and cultural kinship" does not predict the size of gaps or overlaps (Stivers et al. 2009: 10590). Thus, there seem to be cross-linguistic and cross-cultural universals that underlie the timing of interactional conversations.

Contrary to turn-taking, the grammatical structures of languages do not converge on universals, which points to a certain degree of modularity in turn-taking and, possibly, distinct evolutionary heritage (Levinson 2016: 9; Stivers et al. 2009: 10591). Sign languages also seem to have a turn-taking system that corresponds to spoken languages (Levinson 2016: 10). Moreover, turn-taking characterizes the vocal and gestural exchanges of many primates and is documented early in ontogeny, long before language acquisition (Levinson 2016: 10-11; see also section 2.4.1.). Turn-taking exchanges between 2-month-old infants and their adult caretakers indicate that infants engage actively in social interactions (Gratier et al. 2015: 1-2). The infants' ability to engage in early vocal interactions is crucial to their social development, and infants have exhibited flexibility in adjusting the timing of their vocalizations during turn-taking interactions (Masataka 2009: 13). These 'protoconversations' function as the precursors of linguistic exchanges and prepare the ground for "the crucial socio-cognitive skills that precede and enable language use" (Gratier et al. 2015: 9). Acoustic features such as intonation contours, pitch range, and rhythmic speech appear to shape the temporal organization of the turn-taking interactions (Gratier et al. 2015; Levinson 2016). These vocal interactions between adults and infants seem to be mutually regulated, which suggests that infants have a precocious ability of perceiving "contingent relations" and "timing in social interaction" (Gratier et al. 2015: 2).

All the above suggest that turn-taking has old evolutionary roots that precede language phylogenetically,[15] whereas linguistic structures and linguistic diversity are

[15] *Phylogeny* describes the evolutionary development of a species, or of a particular trait of an organism, whereas *ontogeny* pertains to the developmental history of an individual organism, or of a trait during an organism's lifetime.

probably the product of cultural evolution (Levinson 2016: 12). Human conversations are well-coordinated and this stems from the speakers' prosodic modulations, which are the source of cues that regulate turn-taking by aiding speech processing (Filippi 2016: 2-3). The universality of turn-taking patterns and the implications to language evolution show that language processing and production should not be studied separately from dialog (Levinson 2016: 6). The pressure for quick responses during conversational exchanges prioritizes information compression and the search for systematic rules of interpretation (Levinson 2016: 9). This links to pragmatic communication, which has been largely ignored by language evolution studies, but insights from pragmatics are far-reaching and may shed light on the origins of language (Scott-Phillips: 2017).

2.2.3. Experimental evidence

Filippi, Gingras, and Fitch (2014) tested the effect of pitch prominence on word learning in the laboratory. They made an artificial language consisting of nonsense monosyllabic words. They chose three target words which were associated with a semantic category (/ga/ for animals; /lu/ for mountains; and /mi/ for humans) and combined them with other random monosyllabic words into five-word utterances. Each of these utterances was presented over headphones to the participants and each target word consistently occurred with an image corresponding to its meaning. For example, the utterances "su**ga**fokezi" and "**ga**puladero" were paired with animal images (Filippi, Gingras & Fitch 2014: 2). The utterances were presented in six different conditions: (1) simple co-occurrence between target word and image; (2) target word sounding at a higher pitch; (3) random word sounding at a higher pitch; (4) double duration of the target word; (5) extraneous acoustic cue associated with the target word (buzz cue); and (6) extraneous visual cue associated with the target word (change in screen color) (Filippi, Gingras & Fitch 2014: 2-3).

The results revealed that only the condition of the exaggerated pitch of the target word significantly aided word learning among the participants. The pitch deviation was higher by an octave, resembling infant-directed speech (IDS) (Filippi, Gingras & Fitch 2014: 3). The results suggest that prosodic cues are more important in word learning than lengthened duration, visual cues, or other acoustic cues. The participants of this experiment had to identify the target word and infer its meaning.

Pitch prominence was critical in word learning, indicating that the effects of IDS go beyond conveying phonological information or assisting word segmentation (Filippi, Gingras & Fitch 2014: 1). The study by Filippi, Gingras and Fitch provides evidence that statistical learning is not a sufficient factor in word learning, but when it is combined with prosodic cues word learning is greatly enhanced.

Filippi et al. (2017) conducted a study contrasting prosody, verbal content, and facial expression in an emotion identification task. Aiming "to address the relative saliency of multiple communication channels", they used synonyms of *happy* and *sad* which were spoken with a happy/sad prosody and a happy/sad facial expression (Filippi et al. 2017: 881). The participants listened to the words and tried to identify the emotions. They had two response options ('happy' or 'sad'), and their responses were measured in terms of their reaction time and accuracy. The results show that facial expressions were processed faster than prosody, which was in turn processed faster than verbal content. Crucially, however, it was prosody alone that interfered with accuracy in emotion identification. When prosody was incongruent with verbal content and facial expression the percentage of correct answers dropped significantly; in contrast, facial expressions or verbal content alone were not sufficient to bias emotion identification (Filippi et al. 2017: 885-886). This suggests that prosody biases the processing of verbal content, at least in affective words (Filippi et al. 2017: 886-887). The dominance of prosody over verbal content in communicating emotions is in line with the view that communication based on prosodic modulations is evolutionarily more ancient than segmented speech (Filippi et al. 2017: 888).

In another experimental study, Nygaard, Herold, and Namy (2009: 130) investigated whether different speakers make use of "reliable and consistent prosodic cues to meaning", and whether these cues are recognized across listeners in order to derive word meaning. Nygaard, Herold and Namy (2009: 131) asked a set of speakers to employ IDS in novel antonym adjectives, incorporating them in simple questions. By way of illustration, the question "Can you get the *blicket* one?" was produced thrice: in one case *blicket* was pronounced with neutral prosody, in another it was pronounced to mean *big*, whereas in another it meant *small* (Nygaard, Herold & Namy 2009: 131-135). The questions were then presented to a group of listeners, who were instructed to choose one of the two antonyms as a possible interpretation. The results indicate that

listeners inferred the intended meaning from prosody alone. However, when the speakers' intended word meanings were mismatched with the adjective pairs that were offered as possible choices to the listeners, "performance was significantly worse than when [...] choices matched" (Nygaard, Herold & Namy 2009: 140). For example, when listeners had to select *happy/sad* when the speaker's intended meaning was *tall/short*, they did not perform well.

These findings indicate that listeners inferred meaning from "prosodic cues that specifically reflected particular conceptual domains" (Nygaard, Herold & Namy 2009: 139). That these prosodic cues were consistent across speakers provides compelling support to the view that prosody is essential in understanding word meaning, in addition to understanding the structural and affective components of language (Nygaard, Herold & Namy 2009: 140-141). Therefore, prosody may reinforce semantic interpretation beyond the domain of emotions and beyond the arbitrariness of form-meaning mapping. Nygaard, Herold, and Namy's work shakes established beliefs on how meaning is conveyed and challenges the traditional dichotomy between linguistic and nonlinguistic properties of language. It also indicates that prosody may be "an evolutionarily preserved property that is integrally tied to the communicative success of language" (Nygaard, Herold & Namy 2009: 142).

2.3. Developmental evidence

2.3.1. Innate predispositions to language and music

A wide array of evidence suggests that the capacity for language evolved through natural selection (see also section 5.2). An important indication of this in infants is babbling. Prelinguistic, 7-month-olds (even deaf infants) start babbling spontaneously, without imitating adult speech (Patel 2008: 359). After 8 months, their babbling resembles more and more their native language (de Boer 2005: 284) and the first signs of the capacity for vocal learning appear soon after (most babies utter their first words at around 12 months of age [Tomasello 2003: 36]). The ability to produce more complex linguistic structures and speech sounds develops by the age of 3 or 4, preceding other motor skills (Patel 2008: 362). This critical period for language acquisition is shared with sign languages, which are structurally and expressively comparable to spoken languages

(Patel 2008: 362-364). Additionally, the obvious disadvantage of failure to acquire language in any human society (as opposed to music) points to the direct role of natural selection in shaping the evolution of language (Patel 2008: 366). In contrast, for some theorists, music is not considered a biological adaptation, but merely a cultural construct (e.g., Pinker 1997: 529-539).

Such views are typically characterized by narrow definitions of music as being culturally conditioned (cf. Morley 2013: 3-8) and are justified by the apparent absence of musical universals cross-culturally (see Savage et al. 2015). If one thinks of musical capacity as the ability to regulate intensity, pitch, and contour in sound sequences, or the ability to extract information from such sequences (Morley 2013: 8; see section 2.3.2) and treats it as a complex and multi-faceted phenomenon (Fitch 2006: 174), there are plenty of "musically *relevant* traits" (Patel 2008: 377) that appear very early in ontogeny and seem to be biologically innate. Infants, unlike most musically untrained adults, are capable of perceiving "absolute pitch cues"[16] when they process "nonlinguistic tone sequences" (Patel 2008: 47), and they can also "detect pitch changes of a semitone or less, even when the melodic contour is unchanged" (Trehub 2003: 669-670). Moreover, less than 6-month-old infants are sensitive to melodic contour and changes in a melody, and they are also able to discriminate consonant from dissonant musical sequences, showing a preference for the former (Trehub 2003: 669-670; Patel 2008: 373). Infants are also able to identify the same melody played in different pitch ranges, which indicates that the capacity for relative pitch processing may have been favored by natural selection (Patel 2008: 396). This may have its origins in intonation perception during pitch processing, as it would allow one to recognize the same intonation contour regardless of the speaker's voice pitch range (e.g., that of a male adult versus a child) (Patel 2008: 385-386). In addition, prelinguistic infants are sensitive to durational patterning; they are able to perceive rhythmic patterns independent of tempo, based on grouping structure (Patel 2008: 377, 406).

Infants babble for a new months before they utter their first words (Falk 2004: 486). They babble spontaneously but they may also try to imitate the maternal

[16] *Absolute pitch* refers to the ability to detect the absolute frequency of a tone without using its relation to neighboring tones as reference (Patel 2008: 46-47).

vocalizations' pitch contours (Masataka 2009: 15). After the first three months of age, infants' oral-motor control develops rapidly allowing them "to produce vocalizations with distinctively different types of pitch contour" (Masataka 2009: 15). As such, the early vocalizations of preverbal infants are characterized by a musical quality; the flexibility in pitch contour production emerges early, preceding the articulation of phonological sounds.

The musical aspects of language (intonation, rhythm, stress patterns) are perceived by infants during the first months of their life, whereas strictly linguistic or less musical features (referential meaning, syntax) are learned later (Brandt, Gebrian & Slevc 2012: 6). In fact, the auditory experience begins in utero; at least two months before birth, fetuses can hear and learn (Houston 2016: 47). Fetuses can detect different pitches, rhythm, and vowels, but not consonants (Brandt, Gebrian & Slevc 2012: 8; see also Patel 2008: 383). Furthermore, infants prefer IDS to emotionally neutral adult-directed speech (ADS), presumably because IDS conveys a more positive affect through its exaggerated prosodic patterns, slower speed, and greater pitch variation (Houston 2016: 49). They also seem to prefer ID singing than ID speech, apparently for the same reasons: singing seems to be emotionally more expressive than speaking (Trehub 2003: 671; Patel 2008: 381-382). This seemingly inborn preference for emotive sounds might be rooted in the social nature of human interactions, especially musical behaviors, where sharing feelings is primary (Trehub 2003: 672).

In addition, that language perception and acquisition appear to be more precocious and effortless than musical development may be due to exposure and cultural perceptions of what constitutes music (in Western cultures at least) (Brandt, Gebrian & Slevc 2012: 9-10). The development of musical abilities (even without formal musical training) is no more time-consuming than linguistic development (Brandt, Gebrian & Slevc 2012: 9). Furthermore, compared to adults, neonates show more overlapping brain activations in speech and music stimuli (Kotilahti et al. 2010). Hemispheric differences probably reflect asymmetries in "auditory processing rather than specializations for speech or music" per se, and lateralization more directly related to aspects of music or language emerges over the individual's development (Brandt, Gebrian & Slevc 2012: 11). Research on infant cognition indicates that musically relevant abilities in humans are innate (such as the perception of rhythm, pitch, and timbre), and

prelinguistic infants are able to extract emotional information solely from rhythm and tone (Morley 2013: 312). Our capacity to treat sounds in a creative and imaginative manner seems to be a basic property of human cognition, whereas more nuanced linguistic and musical skills (e.g., syntactic processing or sensitivity to differences in harmony) develop gradually on the way to adulthood (Brandt, Gebrian & Slevc 2012: 9, 12).

2.3.2. Prosodic scaffolding

Developmental research shows that IDS occurs cross-linguistically, and its common prosodic patterns communicate emotional information and facilitate social interaction by capturing the infant's attention (Fernald 1989: 1498). IDS does not only convey affective content; it can also aid language acquisition. Infants process auditory input in a holistic manner, "encod[ing] the prominent melodic patterns of IDS and recogniz[ing] these characteristic melodies across variations in speaker" (Fernald 1989: 1508). Prelinguistic infants find IDS "more meaningful than ADS" because the "salient prosodic cues" of IDS give information about the speaker's communicative intent (Fernald 1989: 1508). The communicative effectiveness of IDS is not restricted to prelinguistic infants though. Fernald (1989) manipulated natural speech samples of IDS and ADS to the extent that they became unintelligible. Then she presented the samples to adult listeners, who could only use "intonation contours and certain prosodic features" to identify the communicative intent of the speaker (approval, attention, prohibition, or comfort) (Fernald 1989: 1499, 1506). Listeners performed significantly better in interpreting the IDS than the ADS samples, which demonstrates that, in the absence of lexical content, prosodic cues are reliable indicators to understanding meaning.

Nazzi, Bertoncini, and Mehler (1998) tested newborns' ability to discriminate different languages. They recorded native speakers from different languages and, like Fernald, filtered the sentences by removing the higher frequencies. Thus, they eliminated phonetic information but preserved the prosodic patterns. They found that neonates from French-speaking families were able to discriminate languages that belong to different rhythmic classes (e.g., English from Japanese), but not languages that belong to the same rhythmic class (e.g., English from Dutch). The traditional classification by phonologists to stress-timed, syllable-timed, and mora-timed languages

has not been empirically corroborated; no isochronous intervals have been found in any language between syllables, morae or between stressed syllables (Nazzi, Bertoncini & Mehler 1998: 757). Nevertheless, the notion of rhythmic classes in languages has persisted, indicating that it has "capture[d] some fundamental fact about rhythmic patterns across languages" (Beckman 1993: 458). The data from Nazzi, Bertoncini and Mehler (1998) show that neonates have the capacity to discriminate utterances on the basis of rhythmic information. Such sensitivity to rhythm soon after birth points to its significance in language acquisition and, possibly, to its evolutionary relevance.

Ramus (2002) conducted a series of similar experiments, testing newborns' language discrimination abilities. He found that in (a) natural utterances by different speakers, newborns failed to discriminate between languages that belong to different rhythmic classes (Dutch and Japanese), suggesting that phonetic information may impede on newborns' discrimination abilities (Ramus 2002: 3-6). In (b) resynthesized speech that preserved rhythm and intonation contours, and only some phonotactic information he obtained positive results in line with Nazzi, Bertoncini and Mehler's (1998) findings (Ramus 2002: 6-7). Furthermore, in (c) filtered stimuli that contained only the languages' rhythmic structures (intonation contours were removed) the newborns failed to distinguish the two languages (Ramus 2002: 7-9), whereas in (d) filtered speech that preserved rhythm and only some phonotactic cues (again, intonation was eliminated) the babies apparently discriminated between the two languages, although the effect was weak (Ramus 2002: 7-11). These results suggest that newborns rely heavily on prosodic patterns when they process auditory input. Although rhythm seems to be more salient than intonation, it cannot be effectively isolated (Ramus 2002: 9). The results were more robust when rhythm and intonation were integrated (Ramus 2002: 12). Thus, since infants are not able to make phonemic distinctions at such an early stage of their life (Ramus 2002: 12), prosody is the most important tool they have in processing linguistic input.

Prosody also facilitates the processing of clausal segments. By the age of 6 months, infants are able to perceive prosodic cues (intonation, pauses, stressed syllables) that mark clause boundaries (Houston 2016: 54). Infants use prosodic markers to parse phrase-level units before they start segmenting words (Soderstrom et al. 2003: 264). Soderstrom et al. (2003) provide experimental evidence in favor of a prosodic

bootstrapping in the acquisition of syntax. Infants as young as 6 months old showed sensitivity to prosodic cues to NP and VP boundaries (Soderstrom et al. 2003: 264-265). The infants' preference for syntactic units over non-units is attributed to intonation, preboundary lengthening (i.e., the lengthening of a syllable's vowel), and high pitch excursions in the well-formed NP and VP sequences (Soderstrom et al. 2003: 254-255). Soderstrom et al. (2003: 253) tested the infants' ability to disambiguate syntactic structures only, and not to identify phonemic contrasts (there were not any) or extract meaning. Their work demonstrates that 6-month-olds can successfully use phrase-level prosodic information to process syntactic units.

During the first months of their life, prelinguistic infants process language holistically. Infants have the propensity to imitate maternal vocalizations at the suprasegmental level: pitch contour has multiple communicative functions and its early discrimination and production scaffolds the mapping of meaning onto sound before the acquisition of lexical items (Masataka 2009: 15). Segmental processing and sensitivity to phonemic contrasts emerge later, at around 8-9 months of age (Houston 2016: 51). As infants become more familiar with their native language, they also become more sensitive to its phonology and the sounds that are linguistically relevant. Saffran, Aslin, and Newport (1996) have shown that 8-month-olds use statistical cues to segment continuous speech and locate word boundaries. However, this does not mean that the role of prosody ends there. Infants also use statistical learning to identify prosodic boundaries in clauses and phrases, as well as to infer the rhythmic properties of words (Houston 2016: 54-57). Johnson and Jusczyk (2001: 563) argue, based on their experimental work, that 8-month-olds "rel[y] more heavily on stress cues than transitional probabilities in segmenting words from the speech stream". When stress cues conflict with statistical or phonotactic cues, infants prioritize stress (Johnson & Jusczyk 2001: 564).

The findings reported in this section show that prosody has a key role in language acquisition. From their very first days, infants are responsive to the rhythmic patterns of speech and use prosodic cues as reliable indicators to meaning. Then they deploy prosodic information to process clauses and phrases and, as their awareness of phonemic combinatoriality grows, they start processing linguistic input at word level in addition to utterance level. Thus, the prosodic scaffolding to language acquisition starts

with larger units (clauses and phrases) before reaching the word level (Soderstrom et al. 2003: 264). Infants do not simply respond to the affective content of speech: prosody scaffolds their semantic and syntactic development before the onset of lexical knowledge (see Brandt, Gebrian & Slevc 2012: 6). As the infants' cognitive abilities gradually develop, other learning mechanisms are called upon as well, such as associative learning, recognition memory, and statistical learning (Houston 2016: 52-53). Nevertheless, the importance of prosody in language acquisition and speech perception is evident until adulthood (Fernald 1989), which indicates that it is deeply rooted in human cognition.

2.4. Cross-species comparative data

A cross-species comparison can yield insights into the evolution of language, music, and other aspects of human cognition. Animals across different classes communicate through chorusing, duetting, or antiphonal vocalizations (i.e., alternating turns) in the context of territorial defense and advertisement, group cohesion and coordination, courtship displays, pair-bonding, and parental care (Filippi 2016: 3-8). A key function in these communicative contexts is the modulation of prosodic properties of the vocal signal, that is, manipulation of timbre, pitch, duration, or intensity (Filippi 2016: 3). In the absence of fossilization of hominin vocal organs and other soft tissue, the comparative method provides the framework for inferring the function and structure of organs in extinct hominins and tracing the cognitive abilities of our ancestors (see Fitch 2000: 263, 2006: 181). Shared traits between species are typically identified as either homologous or analogous. Homologies are traits that are shared by virtue of common descent between species and provide valuable information on the phylogenetic level of evolution (Filippi 2016: 3; Fitch 2006: 181). Homologies are critical for deducing "cognitive, neural, and genetic history" on the basis of shared characteristics by related species (Fitch 2017: 5). Analogies, on the other hand, are similar traits that have evolved independently in different lineages and thus did not exist in the common ancestor (Fitch 2006: 181). Analogous traits are the outcome of convergent evolution in phylogenetically distant species, and they evolve as adaptations in response to certain selection pressures (Fitch 2000: 263). In other words, homologies inform us about

ancestry, whereas analogies allow us to assess the possibility of the evolution of specific adaptive functions.

2.4.1. Nonhuman primates

A comparison between humans and nonhuman primates may shed light on evolutionary homologies and, in the case of distantly related primates, strong analogies. Humans are very adept at manipulating formants (see section 2.1.1), which are crucial in articulation. Humans are much more skilled than other primates at controlling the resonances of the vocal tract, something that is made clear by whispered speech (Fitch 2000: 260). Whispered speech is intelligible, even though the vocal folds in the larynx do not vibrate and no pitch is produced; the use of formants in the vocal tract suffices (Fitch 2000: 260). Besides the cognitive abilities that allow humans to shape their vocal tract in a flexible manner, vocal tract anatomy is also important. The human larynx is located lower in the throat than in other apes, which enables us to produce a "wider range of formant patterns" and "[expand] our phonetic repertoire", because we can move our tongue more freely in the vocal tract (Fitch 2000: 260-261). In order to flexibly shape the vocal tract, humans have precise motor control over articulatory parts such as tongue, lips, and jaw (Fitch 2000: 261; Pisanski et al. 2016: 308). An equally far-reaching difference between humans and nonhuman primates is vocal learning. Unlike other primates, humans are capable of acquiring enormous vocal repertoires through imitation (Fitch 2000: 261; Pisanski et al. 2016: 308). Our capacity to imitate novel sounds and use them creatively and voluntarily in novel situations is unparalleled in the animal kingdom.

Despite the fact that they lack articulated speech, several nonhuman primates have the capacity "to modulate the prosodic features" of their vocalizations depending on context (e.g., amount of food source) and their level of arousal (Filippi 2016: 5). Even though voice modulation in these cases is not intentional and under voluntary control, the receivers can perceive the level of intensity of a particular vocal signal and what it conveys (e.g., warning, threat, joy), "behaving in the most adaptive way" (Filippi 2016: 5). Although the vocal repertoires of nonhuman primates are innate rather than learned, the modulation of vocal signals serves communicative purposes (Filippi 2016: 5).

Findings from nonhuman primates that are not closely related to humans may not directly inform us about evolutionary precursors to language and music, but they

allow us to assess the possibility of certain selective pressures and identify vocal behaviors that could be relevant to the emergence of human speech. Several monkey and ape species have been reported to use antiphonal vocalizations flexibly in order to communicate with conspecifics, apparently guided by interactive dynamics akin to turn-taking principles in human communication (see Filippi 2016: 5; cf. section 2.2.2). Gibbons, the lesser apes (see Fig.2), with their "more or less precisely timed vocal interactions", are a case in point (Geissmann 2000: 105). Gibbons, typically mating partners, produce species-specific, song-like vocalizations that are not learned (Geissmann 2000: 109-110). Even though gibbon calls are not learned, they follow recognizable patterns that are combined in duet singing. These vocalizations become progressively more complex and are characterized by well-timed interactions between males and females (Geissmann 2000: 108). Gelada baboons (Old World monkeys) live in unusually large social groups and possess a rich repertoire of complex vocal signals that have deeply social functions (Richman 1993; Dunbar 2016: 208-210; see section 3.1.3). Furthermore, Masataka (2009: 13) reports that adult individuals of the Japanese macaque, a monkey species, engage in vocal exchanges that are characterized by "intercall intervals [...] similar to those obtained in human mother-infant dyad" (cf. section 2.3). These adult monkeys seem to use pitch in a consistently different manner than young individuals, implying that there is some form of learning (presumably through social experience) which could enhance the communicative potential of the vocal signals (Masataka 2009: 13-14).

Findings from the great apes, especially from chimpanzees and bonobos, our closest living relatives (Fig. 2), are the most pertinent to the search of evolutionary homologies. Captive chimpanzees have been observed to produce "attention-getting" vocalizations that are context-specific and have not been documented in wild populations (Hopkins, Taglialatela & Leavens 2007: 281-282). In the presence of unreachable food and a human that could potentially help, chimpanzees produced spontaneous, "species-atypical" calls that were different from the typical food calls (Hopkins, Taglialatela & Leavens 2007: 281). Such vocal signals have also been reported in situations where a human held an object necessary for a problem-solving task unrelated to food (Russel et al. 2005). The fact that these calls were produced in the absence of any explicit training indicates that chimpanzees produced them intentionally, "as a means of manipulating the attentional status of a human" (Hopkins, Taglialatela &

Leavens 2007: 284). Thus, depending on the context, chimpanzees show some flexibility in their communicative vocalizations, which suggests that they have a certain degree of control over their vocal output and a latent capacity to produce novel sounds (Hopkins, Taglialatela & Leavens 2007: 284-285).

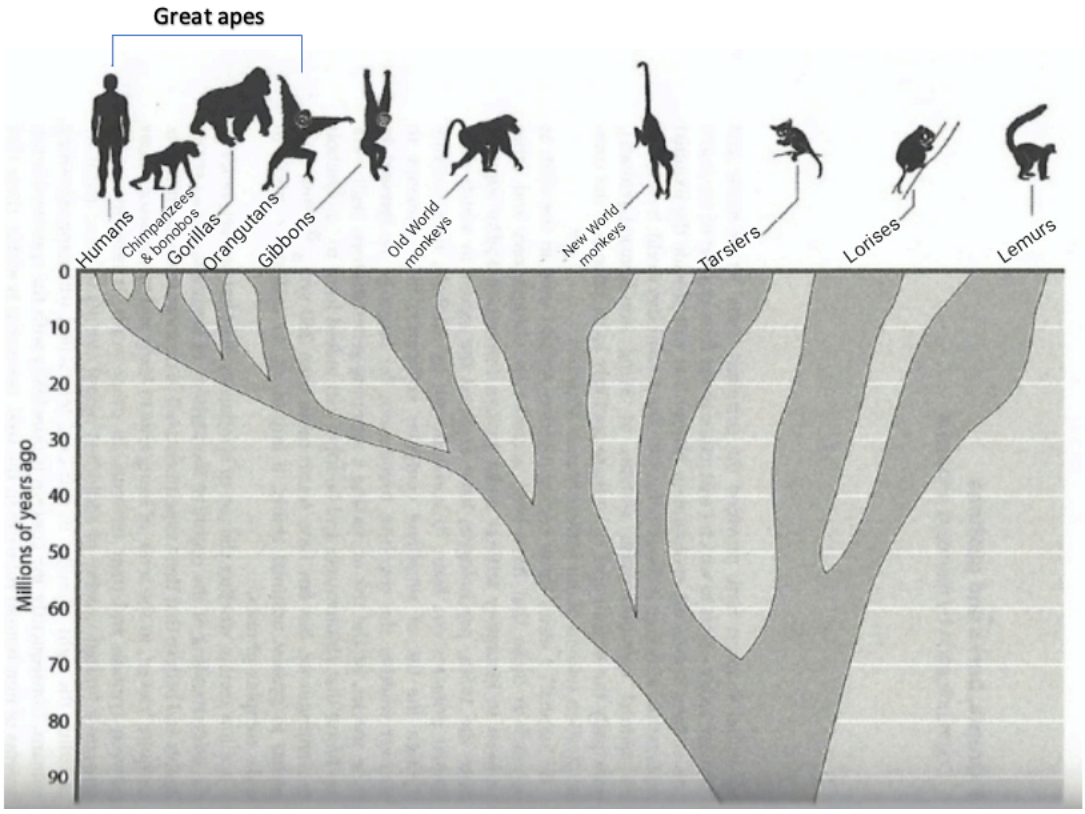

Figure 2. Evolutionary relationships between extant primates. The branching tree shows the last common ancestor and the subsequent splitting of lineages. Humans are genetically most closely related to chimpanzees and bonobos, having diverged from the shared ancestor approximately 6 million years ago (Adapted from Mithen 2005: 106).

Recent research on wild bonobos also reveals greater flexibility in their communicative behavior than previously assumed. As in other primates, much of bonobo vocal signaling is fixed, arising in response to particular stimuli, such as "barking during alarm and laughing during play" (Clay, Archbold & Zuberbuhler 2015: 13). However, bonobos also produce the "peep", a specific high-frequency vocalization, in variegated ways (Clay, Archbold & Zuberbuhler 2015: 1). Clay, Archbold and Zuberbuhler (2015: 9) point out that bonobos produce peep calls across a range of behavioral situations of positive, neutral, and negative valence. Peeps associated with positive valence states are distinctly different from those produced during negative contexts, showing that bonobos exhibit functional flexibility in this type of call (Clay, Archbold &

Zuberbuhler 2015: 11). The way bonobos use their peep calls is reminiscent of the vocalizations of preverbal human infants, who, being still unable to utter words, modulate their voice to communicate affective meaning.

Nonhuman primates are able to coordinate their calls and modulate their voice to produce recognizable patterns that may be contingent to social circumstances (see Filippi 2016: 6). Across different primate genera, females are as vocal as males (Geissmann 2000: 111), which indicates that mate attraction is not a primary function of primate calls. Instead, these calls are thought to be related mainly to within- and between-group communication, regulation of social relationships, territorial advertisement, and pair-bonding (Geissmann 2000: 110-116; Filippi 2016: 6). Furthermore, findings from great apes suggest that their vocalizations are not as fixed as traditionally thought. Chimpanzees have exhibited the ability to produce novel, context-dependent vocalizations in order to manipulate humans' behavior (Russel et al. 2005; Hopkins, Taglialatela & Leavens 2007), and bonobos have shown flexibility in their vocal signals in order to convey emotive meaning in a range of circumstances (Clay, Archbold & Zuberbuhler 2015). Thus, even though apes are not capable of vocal learning and do not possess the anatomical features necessary for human speech, they have a latent capacity for volitional vocal control and functionally driven novelty in their vocal signals (Pisanski et al. 2016: 308-310; see also Masataka 2009: 18-19). This could be interpreted as an intermediate stage between the generative nature of human language and the fixed vocal signals that nonhuman primates typically produce (Clay, Archbold & Zuberbuhler 2015: 14).

2.4.2. Birds and non-primate mammals

Analogies between humans and phylogenetically distant species do not give direct information regarding the evolution of language or music in the human lineage but allow us to evaluate evolutionary theories and the likelihood of selective pressures that led to the emergence of language and music. Although no other animal that possesses an equivalent of human language or music has been found, there are certain components of language and music that are observed in animals unrelated to humans.

Vocal learning, which is different from innate calls and is a prerequisite for cultural transmission, has evolved independently in aquatic mammals (cetaceans and

pinnipeds), bats, and many species of bird (Marler 2000; Payne 2000; Slater 2000; Fitch 2013: 494; Patel 2014: 3-4). These animals are capable of learning long and complex vocalizations (whales are known to even improvise on their learned songs), which are typically produced by males and function to attract mates and repel rivals (Payne 2000: 135; Slater 2000: 49). Nevertheless, there is evidence that female birds have also the capacity to sing (Fitch 2006: 186), and that the vocal signals of songbirds and whales have a social dimension, facilitating interactive coordination and, in some cases, group cohesion (Filippi 2016: 6-7). What these learned vocalizations have in common with the innate calls of these species, as well as nonhuman primates' innate calls, is that they are affective in character, and they lack symbolism (Marler 2000: 40-42; see also Filippi 2016: 5-8). An interesting fact about birds is that, although they have a larynx, they do not use it to make sounds; they use the syrinx instead, their vocal organ (Fitch 2006: 186). The syrinx of songbirds is "arguably the most sophisticated vocal instrument known" (Fitch 2006: 186), and some songbirds have "an individual repertoire of over 1,000 distinct songs" (Marler 2000: 40). This shows that, given the selective pressures for a certain function, animals evolve the necessary adaptations regardless of how complex they might appear to be.

The capacity for audiomotor entrainment, i.e., the ability to synchronize motor actions to an external pulse, is a rare trait that humans share with only a few species (Patel 2014: 1; Kotz, Ravignani & Fitch 2018: 897). This ability to perceive and synchronize to a beat is essential for human music and dance, and it may also have implications for speech prosody. Humans are able to synchronize to a beat (a) in varying tempi, (b) cross-modally (e.g., by clapping, vocalizing, or head bobbing), and (c) in a predictive manner: they create "a mental model of time, rather than simply [...] [reacting] to each stimulus" (Patel 2014: 2). Some species of parrot and a sea lion have been reported to synchronize to a beat in the way humans do, although this does not happen in the wild; the parrots were pets and probably mimicked human behavior, whereas the sea lion received intense training (Patel 2014: 3-4; Kotz, Ravignani & Fitch 2018: 903). Parrots are known vocal learners and sea lions, while not vocal learners themselves, are closely related to seals and walruses, which are capable of vocal learning (Patel 2014: 4). This indicates that vocal learning might play a role in the ability for audiomotor entrainment, although there are probably other reasons involved as well (see Patel 2014:

3). Both audiomotor entrainment and vocal learning "[require] tight auditory-motor coupling in the service of an auditory model" (Patel 2014: 3), which suggests that these abilities may be associated with shared neural circuitry.

Great apes do not exhibit audiomotor entrainment, but they are known to drum on buttresses and trees, stomp with their feet, and hit the ground with their hands (Geissmann 2000: 115-116; Fitch 2006: 183; Kotz, Ravignani & Fitch 2018: 903), and some of these activities may include rhythmicity (Fitch 2006: 183). Moreover, chimpanzees may have "the capacity for anticipatory synchronization" in limited contexts and without tempo flexibility (Patel 2014: 2). This suggests that, like vocal control, they may have a latent capacity for basic rhythm perception (Kotz, Ravignani & Fitch 2018: 907). It is not straightforward whether audiomotor entrainment has any clear adaptive value. Rhythmic singing and dancing in humans have an affective component that greatly facilitates social bonding (Kirschner & Tomasello 2010), and parrots, which have the capacity for both vocal learning and audiomotor entrainment, are "deeply social creatures" (Patel 2014: 4). Therefore, it is possible that vocal learning and sociality are two prerequisites for the evolution of this unusual ability.

The findings discussed so far indicate that language and music have more in common than typically assumed. They both require sophisticated oral-motor learning skills in order to be culturally transmitted. Importantly, elaborate vocal control is a prerequisite for both traits, as it allows individuals to volitionally manipulate pitch, timbre, amplitude, and duration in vocalizations. Furthermore, timbre, rhythm, syntactic structures, and pitch contours are not domain-specific attributes; instead, they are fundamental to both language and music.

Neuroscientific data show that language and music processing and production activate multiple areas of the brain, and that there are often overlaps in brain activity during linguistic and musical tasks. This suggests that (a) despite lateralization tendencies, the two traits cannot be limited to individual cerebral regions, and that (b) there are firm neurological links between language and music. Moreover, it has been shown that pitch contour processing is implicated not only in music but also in speech perception. There is also evidence for cross-cultural and crosslinguistic universals in the perception of temporal patterns. Additionally, experimental findings indicate that

prosodic cues greatly enhance word learning and semantic interpretation, which points to the ancient origins of prosody in successful communication.

Human neonates are biologically predisposed to acquire language. They are also sensitive to changes in melodic contour, and they have the capacity for relative pitch processing. Their babbling, which displays flexibility in pitch contour, anticipates the production of lexical items. Hence, musically relevant features appear to have been targeted by natural selection. In addition, prelinguistic infants are responsive to intonation and rhythmical patterns of speech. They use prosodic information to segment speech and map meaning onto sound; the prosodic cues scaffold their language acquisition.

Cross-species comparative data suggest that the ability for audiomotor entrainment and vocal learning may be associated with shared neural substrates. Thus, any theory of language evolution that addresses the ability for vocal learning should not overlook its relation to audiomotor entrainment. Turning to nonhuman primates, research on various species indicates that their vocalizations can be well-timed, complex, and varied in pitch range. More relevantly, chimpanzees have been observed to produce intentional vocalizations in order to manipulate humans' attention, and bonobos produce vocalizations in diverse ways and across a range of behavioral situations. These suggest that our closest living relatives may have a latent capacity for vocal control. This could allude to the nature of vocal exchanges among hominins after the split with our shared ancestor with chimpanzees and bonobos.

3. Musical protolanguage proposals

In this section, a sample of musical protolanguage proposals that have been made in the literature is presented. There is no agreement on the nature of the putative protolanguage. Some authors emphasize the prosodic properties of speech in reconstructing the evolutionary history of language, whereas others posit that music-making and singing have ancient origins, preceding the emergence of articulated speech. In the following, an attempt is made to evaluate the plausibility of the proposals. The focus is on (a) the selective pressures and mechanisms assumed to have been relevant for the emergence of language, and (b) the assumptions that are made about the type

of protolanguage that preceded the emergence of the fully developed language attested in modern *Homo sapiens*.

3.1. Selective pressures

3.1.1. Sexual selection

Sexual selection is about reproductive success that directly depends on mate choice and mate competition (see Miller 2001: 8; Mithen 2005: 176-177). Evolutionary change through sexual selection occurs "due to heritable differences in the ability to attract sexual partners [...] [or] repel sexual rivals" (Miller 2001: 445). Sexually selected traits are conspicuous, and they evolve either because they attract the opposite sex or because they grant increased prowess to someone who competes with members of the same sex for access to mating partners (Ayala 2019). Mate choice determines the appearance and behavioral traits the offspring will inherit (Mithen 2005: 176-177). For example, if females acquire a preference for a male trait, this trait may be elaborated simply because the females favor it. Males that have the desired trait will have more offspring that will inherit the trait (if they are males) or the preference for it (if they are females). This may then select for individuals that have a more pronounced desired trait or a stronger preference for it. The notion that language or music evolved through sexual selection means that they evolved in the service of vocal competition, for territorial defense and advertisement, or as courtship displays (see Fitch 2010: 490).

Sexual selection has the virtue of explaining the "rapid evolution of unusual traits" through the *runaway* process (Fitch 2010: 490): runaway sexual selection can explain traits that are "striking", "costly", "attractive to the opposite sex", and "have little apparent survival value" (Miller 2001: 97). Sexual selection can potentially explain almost any odd trait, which makes it a very powerful theory. However, there are some constraints. It produces differences between the sexes (the peacock's tale is a widely cited example) and typically demands polygamy (see Miller 2001: 74-76). Furthermore, sexually selected traits emerge late in ontogeny, at sexual maturity (Fitch 2010: 490). These facts pose a major obstacle in considering sexual selection as the main driver of the evolution of human traits such as language and music.

Darwin (1871) was the first to hypothesize that human language and music might share a common precursor. He observed the chirping of songbirds and noticed that their songs are not instinctive; like human language, they are learned and culturally transmitted (Darwin 1871: 55). Darwin (1871: 55-56) draws a parallel between infant babbling and the imperfect song of young birds, and he also points to the different birdsong "provincial dialects" of the same species. While he acknowledges that language requires complex vocal learning, he distinguishes it from other animal vocalizations, stressing that the capacity for language depends on the mental abilities of the human brain (Darwin 1871: 54). Nevertheless, the analogy between human language and birdsong was probably what led him to advance the idea of a musical protolanguage for some early ancestors of humans. Contemporary research has confirmed that many species of bird are capable of learning long and complex vocalizations (Slater 2000: 50), and that there are "behavioural, neural and genetic similarities between auditory-vocal learning in birds and human infants" (Berwick et al. 2011: 113). Along with the fact that vocal learning has evolved in mammals as well, this has led Fitch (2013: 494-495) to conclude that convergent evolution provides strong "empirical basis for estimating the likelihood" of the evolutionary emergence of language (see also section 2.4).

Darwin (1871: 56-57) postulates that "the mental powers" of our early ancestors improved somehow and this enabled them to imitate natural and animal sounds, and human voices. He believes that this led to the emergence of a protolanguage that was truly musical, and that its emergence was primarily driven by sexual selection: drawing on the bird analogy, he proposes that its principal functions were to attract mates and to oust rivals. Indeed, the songs of birds and some other animals, such as whales, seem to have exactly these functions during the breeding season (Slater 2000: 59; Payne 2000: 135), but this evidence in and of itself is insufficient to conclude that sexual selection was indeed the driving force of language evolution. Human language occurs in a much broader context than these animal vocal signals and is under our volitional control. Moreover, unlike human language, birdsong does not exhibit nested dependencies, is restricted in length or structure, and lacks symbolic semantics (Berwick et al. 2011: 119-120). Thus, it lacks the "lexical syntax" that defines human language (Marler 2000: 36-37). According to Marler (2000:41), birds (and whales) have "phonological syntax": they generate novel sound sequences, but their individual components do not carry any

particular meaning. Importantly, these studies show that birdsong and other animal learned vocalizations are affective in character and occur mainly in a mating context. As noted in section 2.4, analogies between distantly related species only allow us to evaluate the likelihood that a trait represents an adaptation to specific environmental conditions; unlike homologies, they do not inform us about genetic history. Consequently, one should be cautious not to ascribe too much importance on this analogy (see also Morley 2013: 133).

In an attempt to revive Darwin's theory, Miller (2000, 2001) claims that sexual selection was the chief pressure that shaped music, language and human cognition in general. Because he does not see any direct survival benefits in music or language, he proposes that they originally functioned as courtship displays. According to Miller (2000: 329), most complex vocalizations across species are produced to attract mates and thus this function "should be considered the evolutionary null hypothesis" for human music. Miller (2000: 338-341) sees musical abilities as reliable *fitness* indicators: costly sexual ornaments that confer a reproductive advantage by attracting the opposite sex (see Miller 2001: 102-104). He extends the idea of fitness indicators to human intelligence and language (Miller 2001: 129-134). He further writes that pidgin languages with their structural simplicity and small vocabularies suffice for pragmatic communication and "ordinary survival functions" (2001: 371), and thus human language in all its complexity and vast vocabulary is a fitness indicator of intelligence (2001: 369-375). Because he acknowledges some of the problems that arise with runaway sexual selection and fitness indicators, he invokes mutual choice to account for the equal language skills and intelligence between women and men, and also for the pairbonding in long term relationships (Miller 2001: 375-376; 383-386).

It is startling that Miller (2000: 337) completely disregards the fossil record, claiming that "[i]t is just not very important whether music evolved 200,000 years ago or 2 million years ago, or whether language evolved as a precursor to music". Archeological findings make it possible to trace physiological adaptations to speech (and singing), and also to detect the evolution of relevant genes (see section 5.2). Miller is not concerned with the question of when music and language might have evolved and what their evolutionary trajectory might have been. He also leaves the relationship

between music and language unaddressed. Although he acknowledges the social nature of primates and the social function of language, he makes no reference to hominin social structures, and he sees language mainly as an indicator of social status and intelligence display that was "favored by sexual selection" (Miller 2001: 367-368). His statements underlie his conviction to an adaptationist approach, focusing on the functionality and fitness of human traits. However, he fails to see that music and language are complex cognitive abilities that can be better understood in their broader context and if their underpinning mechanisms are first identified. Curiously, he rejects convergent evolution as a possibility, claiming that sexual selection "never happens the same way twice" (Miller 2001: 20). However, as noted above, convergent evolution allows us to assess the likelihood that a trait has evolved as an adaptation to specific environmental conditions, and thus it has the potential to explain the evolution of similar traits (such as vocal learning) in different species.

Despite insisting on the dominant role of sexual selection, Miller is vague on how the various human adaptive traits emerged in the first place. He writes: "[o]nce the rudiments of language started to evolve, for whatever reason, our sexually motivated ancestors would probably have used their heritable language abilities in courtship" (Miller 2001: 353). In this statement, and in several other parts of his book (e.g., 2001: 10, 148, 221-222), it is made clear that Miller's sexual selection kicks in after the advent of language. Thus, there is a certain degree of circularity in his argumentation. Furthermore, he refers to language and music evolution as if they were monolithic events. Even though he denies that there is a "single gene for language" (Miller 2001: 25), he does not attempt to account for different evolutionary phases, nor does he mention that language or music must have evolved in a piecemeal fashion. By neglecting to do so, Miller does not seem to be fully embracing the multiplicity of language and music's underlying mechanisms, what Fitch (2017: 4) describes as the "multicomponent approach".

In addition, Miller conflates biological evolution with cultural evolution in several of his examples. The biological evolution of a trait refers to neurological specializations that play a direct role in its expression (see Patel 2008: 358; Morley 2013: 194), whereas culturally evolved traits and behaviors do not have a biological basis and are acquired "through teaching, imitation and other forms of social learning" (Laland, Odling-Smee &

Myles 2010: 138). By way of illustration, it is questionable whether morality, religion, or ornate language should be attributed to biological evolution as Miller (2001: 18) implies; they could be simply regarded as cultural inventions. Equally, conspicuous consumerism does not need to be interpreted always as a sexual display; it simply may reflect modern human ideology and culture. In addition, Miller (2001: 369, 390) sees the "excessive" vocabularies of adults with the "many useless synonyms" as fitness indicators. Although there might be a grain of truth in ornate language as a fitness indicator, one should not underestimate the role of the increasing social complexity and the invention of writing in the last few thousand years. Moreover, the fact that agriculture and other major technological advancements of the so-called Neolithic revolution appeared only in the last 10,000 years, long after the emergence of anatomically modern humans, does not mean that large brains did not confer any immediate survival payoffs, as Miller (2001: 17-18) alleges. Changing climatic conditions at the end of the last glacial period were catalytic for many of the developments that Miller attributes to sexual selection.

Regarding the origins of music, Miller (2000) places excessive emphasis on young male courtship displays. However, no other apes are known to use vocalizations to attract mates, and nonhuman primate calls are innate (Marler 2000: 43). Crucially, both men and women appear to have the same proclivity for music and language, which appear very early in human ontogeny, whereas sexually selected traits manifest late, usually with sexual maturity (Fitch 2010: 490-491). The examples of male dominance in music production (e.g., Miller 2000: 331) downplay cultural factors and the values of patriarchal modern society. As for the language skills, some have even argued that women are superior to men (see Fitch 2013: 496-497). Additionally, sexual dimorphism has been decreasing in the hominin lineage (Dunbar 2016: 132), which points to a waning influence of sexual selection.

The theory of runaway sexual selection that Miller (2000: 342-343) invokes produces sex differences in the evolved traits and tends to be very fast. Thus, Miller (2001: 97-98) concedes that "mutual mate choice" is probably a more plausible explanation for the evolution of human language, music, and the brain. But mutual mate choice counters the runaway theory that Miller draws on heavily. It also renders fitness indicators rather irrelevant. So, mutual choice is not what best describes sexual selection processes; if mutual choice played a prominent role in the evolution of human cognition,

this indicates that sexual selection was not a primary evolutionary force. Mutual choice is contingent on social structure and on an advanced *theory of mind*: a form of social cognition that includes mind reading; namely, the ability to recognize that other individuals have intentions, beliefs, and "an informal theory [...] about what having a mind is like" (Dunbar 2016: 43-45). Consequently, mutual choice has other implications that Miller does not discuss; it may inform us about kin relations, reciprocal pair-bonding, (social) monogamy, the role of males in child-rearing, and social complexity (discussed in sections 3.1.2 and 3.1.3).

Merker (2000: 315) speculates that synchronous chorusing among early hominins could have been fundamental in human evolution, launching the split from our last common ancestor (LCA) with chimpanzees. Because humans, among primates, have the unique ability of entrainment to an external regular pulse, Merker posits that our early ancestors developed this ability as a reproductive strategy; males looking for sexual partners joined their voices with the hope of attracting the attention of migrating females. According to Merker, such vocalizations were the forerunner of language and music. Additionally, synchronous chorusing involving vocal learning may have triggered brain expansion (Merker 2000: 320-321). The core of Merker's proposal lies on the fact that in many tribal societies women leave their natal groups and join neighboring ones to get married. Since this phenomenon of "female exogamy" is also observed in our closest relatives, the chimpanzees, Merker (2000: 317-318) hypothesizes that the first hominins shared this habit, which put selective pressures on males that wanted to lure prospective mates. The presumed cooperative chorusing appeared after our LCA with chimpanzees, for there is no evidence of beat synchronization in other primates.

As is the case with Miller, the main problem with Merker's theory is that it concentrates on males. The postulated cooperative male calls do not explain how women acquired the ability for mutual entrainment. To support his thesis, Merker invokes bipedalism. He states that upright posture would have facilitated the coordination of vocalizations by helping to synchronize bodily movements "as a form of dancing display" (Merker 2000: 319-320). Bipedalism would have proved useful indeed, but this does not strengthen Merker's argument. Of course, theories of human evolution should not overlook bipedalism and its consequences, as it is one of the first traits that

distinguished hominins from other apes. However, when we consider the origins of language and music, bipedalism is of little significance if the rest of the arguments do not have solid foundations. Although bipedalism and the anatomical adaptations that come with it must have facilitated the emergence of Merker's putative protolanguage, it is not evidence that a musical protolanguage evolved out of it. Put it otherwise, bipedalism is neither a necessary condition for the emergence of language and music, nor a sufficient one. This may be exemplified by the fact that some animals phylogenetically distant to humans, such as parrots, have the ability for vocal learning and beat synchronization (see Patel 2014: 3-5). Finally, an important problem with attributing human brain expansion to vocal learning (Merker 2000: 320-321) is the large number of other animals of different clades that are capable of learning long and complicated vocal sequences (Fitch 2013: 494). Thus, it is questionable whether vocal learning was the triggering factor for brain growth. In spite of the unconvincing arguments regarding the functional origins of language and music, Merker's proposal raises an interesting issue: that of the relationship between the ability for rhythmic synchronization and vocal learning, and whether the former is contingent on the latter (see also Patel 2008: 402-211; 2014).

In support of the sexual selection theory, Leongomez et al. (2014) adduce experimental evidence where it is shown that both women and men modulate their voice and increase pitch variability in order to be perceived as more attractive to potential mates. Such vocal modulations are detectable by listeners and seem to influence their judgment on prospective mates (Leongomez et al. 2014: 490). In addition, variability in voice pitch, especially among women, is observed in the presence of sexual competitors (Leongomez et al. 2014: 489). Moreover, Pisanski et al. (2016: 308) note that voice timbre and pitch control "may have emerged to exaggerate socially relevant traits such as body size and threat potential even before the advent of articulated speech". The ability of some enculturated apes to exhibit timbre and pitch control indicates that nonhuman primates have a latent capacity for vocal control, but not for vocal learning (Pisanski et al. 2016: 308-310; see also section 2.4.1). Thus, the fact that vocal flexibility is employed in humans and other animals in relation to size exaggeration or sexual cues may point to the role of sexual selection during early hominin vocalizations that served as the

precursor of human language and song (cf. Fitch 2010: 491). Otherwise, Leongomez et al.'s (2014) findings may simply point to the advanced social cognition that humans possess.

The evidence examined so far suggests that sexual selection did not have a role to play in the evolution of language and music, and even if it did it was a minor one. Despite the unidimensional character of Miller's proposal, it succeeds in directing our attention to aspects of human evolution that may have been ignored. If we accept that sexual selection has been relevant, it is probably reasonable to propose that it may have exerted greater influence during the early, prelinguistic and premusical stages of hominin evolution. However, except for the paralinguistic traits discussed here, sexual selection theories have not been able to provide any evolutionary explanations for defining characteristics of music and language, such as syntax or semantics (Oesch 2019: 8; see also Fitch 2013: 495). Human evolution has lasted more than 5 million years and has gone through several distinct phases. Thus, it has arguably been influenced by multiple factors throughout the years. Hence, the various theories on the topic need not be mutually exclusive. Merker's attempt, despite its weaknesses, acknowledges this by trying to bridge the gap between sexual and social selection theories.

3.1.2. Kin Selection

Kin selection refers to the role of blood relatives when assessing an individual's *fitness*: the genetic predisposition to survive and reproduce successfully. Kin selection occurs when an individual is generous towards genetically related ones, improving their overall fitness (Britannica 2018). Kin selection can explain altruistic behavior, because the beneficiaries share the same genes with those that act altruistically. In this way, one's genes may be carried on by boosting the fitness of a relative, even if this comes at the cost of the aid giver. The influence of kin selection processes is proportionate to genetic relatedness: individuals tend to favor more close relatives (Miller 2001: 441). Therefore, parental care (probably the most common form of altruism) can be "readily explained by kin selection [...] because it increases the reproductive success of the parent's genes" (Britannica 2018: 2).

Although the papers discussed in this section focus on mother-infant interactions, kin selection processes in the evolution of language may be implicated not only in parent-offspring interactions, but also in other blood relatives, typically in vocal exchanges between adults and younger kin. Kin selection accounts of the evolutionary emergence of language and music draw on developmental evidence and may help to explain why language and musical abilities appear early in ontogeny (Fitch 2013: 497). Additionally, the slow reproductive rate and the prolonged childhood in humans put pressure on the survivability of offspring, which can potentially explain why language evolved in our lineage and not in other species that exhibit vocal learning (Fitch 2013: 497). Kin selection theories of language and music evolution focus on the affective component of adult-infant interactions. However, affective vocalizations are preponderant in adult-to-adult interactions among nonhuman primates, and thus there is no reason to believe that infant-directed speech (IDS) was evolutionarily more relevant than adult-to-adult interactions (see Morley 2013: 209-210), given the "intense sociality" of humans and chimpanzees, with whom we share our LCA (see Fitch 2010: 240). Nevertheless, kin selection theories need not be mutually exclusive with social selection theories (see section 3.1.3); insights from kin selection models help to build a more plausible scenario of how language may have originated.

In a thought-provoking paper, Falk (2004) sets forth the view that mother-infant interactions in early hominins were a catalyst in the evolution of language. As early hominins became increasingly bipedal, infants lost their ability to ride on their mothers' back, who in turn occupied their hands more frequently in order to forage and carry food, thus losing physical contact with their babies. As mothers were temporarily separated from their babies to gather food, they used their voice to soothe them in a similar way that contemporary human mothers sing lullabies to reassure their babies. Falk (2004: 500) terms this the "putting the baby down" strategy. She (2004: 500-501) argues that maternal prosodic vocalizations provided comfort and lulled distressed infants. These rhythmic vocalizations, which were initially affective in character, gradually became instructive and facilitated socialization, forming the basis for the emergence of protolanguage (Falk 2004: 501-502).

In support of her thesis, Falk mainly draws on data from IDS and compares it to mother-infant communication among nonhuman primates, who apparently lack the ability for human-like ID vocalizations. She illustrates the importance of IDS in providing emotional support to infants. She also argues convincingly that IDS reinforces language acquisition: the typically exaggerated rhythmic and melodic prosodic patterns of IDS attract infants' attention, while meaning and grammar are not processed at the earlier stages (Falk 2004: 496). Moreover, infants start to babble in the prosodic patterns of their language before they utter any words, and they develop a sense of turn-taking through the vocal interactions they have with their mother (Falk 2004: 496; see also section 2.3). All the above evidence stresses the importance of prosody in language acquisition and in human cognition. It is therefore reasonable to propose that prosody could have had a pivotal role in language evolution and that it preceded lexical semantics and syntax.

Motherese, which is directed toward individuals that are in the early stages of their ontogeny, could quite possibly be analogous to early forms of hominin communication, but does this mean that the earliest hominin vocalizations should be limited to mother-infant interaction? Because IDS scaffolds language acquisition, it does not follow that ID vocalizations scaffolded the emergence of protolanguage. In Falk's argument, ontogeny is projected onto phylogeny. The ontogenetic development may reflect evolutionary processes at the phylogenetic level, but this does not signify that mother-infant interactions, which support ontogenetic development, must have driven the evolution of language phylogenetically. Therefore, there is an explanatory gap in using ontogeny as the starting and focal point in order to explain the phylogenetic origins of language. By downplaying male adults or adult-to-adult communication, Falk is rather vague on how mother-infant interactions led to vocalizations in a broader social context. Furthermore, the assumption that juveniles started sharing "simple instructive utterances" (Falk 2004: 503) seems problematic: the ability to instruct and share instructions presupposes that a certain form of protolanguage was already present, so it cannot be a cause for its emergence. Also, language acquisition and other practical skills are normally learned "via observation and imitation" and not through explicit instructions (Tallerman 2013: 480). Hence, there must have also been other selective pressures at play.

Falk indeed mentions that her proposal is consonant with social selection theories, as the "mother-infant dyad is fundamentally social" (Falk 2004: 501). Therefore, the mother-infant interactions that Falk proposes do not have to be envisaged in isolation from the broader social context. Another point of consideration might be the time that the putative mother-infant interactions originated and evolved. Falk (2004: 491) sets her proposal in the period of the "late australopithecines/early *Homo*", which is roughly 2.5 million years ago (MYA). Whereas the amount of data from the more recent *Homo species* allows us to make several plausible inferences about the lifestyle of our ancestors, the truth is that we still know very little about earlier hominins such as the australopithecines (Dunbar 2016: 10). If Falk's (2004: 500) "putting the baby down" hypothesis is correct, the putative ID vocalizations must have been subject to greater selective pressures during these early years of human evolution. However, as it is discussed in sections 3.1.3 and 5.2, the findings converge on the period of the *Homo ergaster/erectus* expansion (ca. 1.8 - 0.5 MYA; see Fig. 3) as the most critical in the evolution of speech and language.

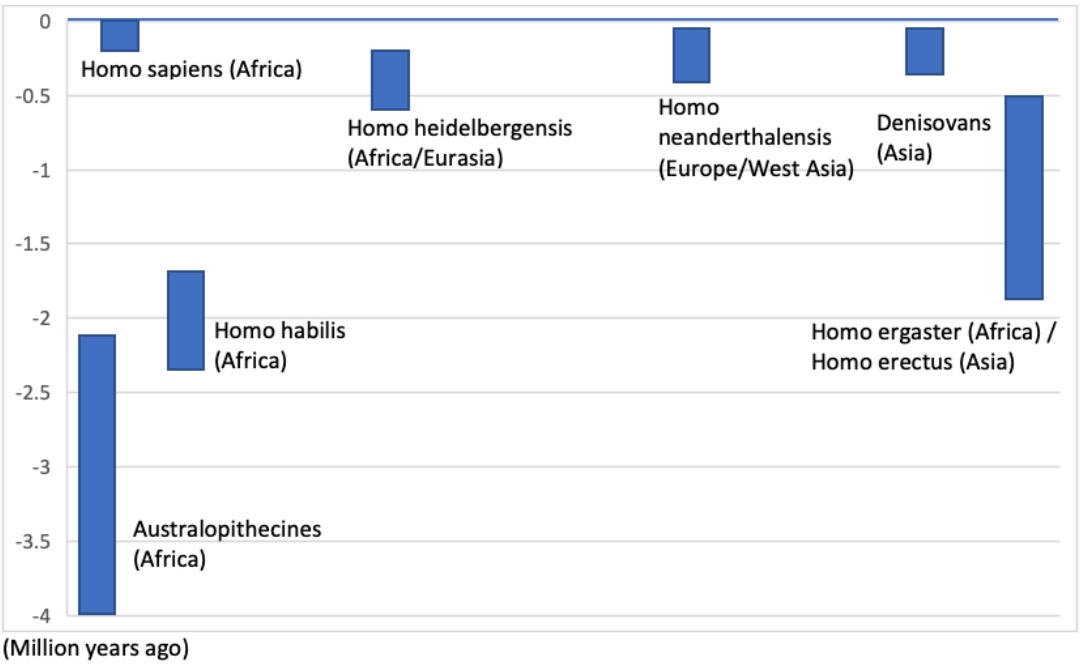

Figure 3. The last 4 million years of human evolution, showing the hominin species mentioned in the text and their approximate time spans. *Homo ergaster* emerged from the australopithecines in Africa and later evolved into *Homo heidelbergensis*. The expansion of *Homo heidelbergensis* in and out of Africa led to the development of *Homo sapiens*, the Neanderthals, and the Denisovans. The Denisovans do not have an agreed species classification name because of their scant fossil record. Drawn on basis of data from Dunbar (2016) and Ko (2016).

Falk gives us a plausible explanation of how bipedalism could have played a role in the evolution of language. It may be the case that the setting of the conjectural mother-infant interactions that Falk describes provided the backdrop for the onset of language evolution. The exaggerated prosodic patterns of IDS have a distinct musical quality that seems to be critical in aiding language acquisition at the early stages. The maternal lullabies and IDS hint at the nature of protolanguage, which may have originally allowed hominins to achieve communicative ends through the musical elements of speech rather than through syntactic structures. As hominins evolved and became less dimorphic, their brain size increased, and their social organization became more complex (see also section 3.1.3). In this context, the tentative ID vocalizations could have been extended to other members of the group, including males. Despite its narrow scope, Falk's proposal earnestly grapples with the functional pressures at the very origins of language evolution. As the author herself acknowledges (Falk 2004: 501), her proposal can be complemented by other theories that focus on the social aspect in hominin evolution.

According to Dissanayake (2000), the origins of music lie in mother-infant interaction. As human neonates were born increasingly premature and helpless, mothers and infants developed affective and coordinated musical interactions as a way of sustaining maternal care. These rhythmic interactions helped to create positive affect and strong emotional bonds between mothers and infants (Dissanayake 2000: 390). Dissanayake differs from Falk in that she does not point to reassuring maternal lullabies, but to ritualized vocal interactions that were accompanied by facial expressions, gestures, and other kinesic behaviors. Dissanayake does not mention when the musical mother-infant interactions might have evolved. Considering that she believes them to be the origins of music (and presumably language) in the biological sense, one could not place them in the very recent evolutionary past. Dissanayake (2000: 394) explicitly argues that mother-infant interactions are literally musical in nature, and she draws on developmental studies that stress the benefits of motherese to infants.

The parallels that Dissanayake draws between music and IDS offer support to the idea of a common precursor of music and language, but they do not necessarily show a causal relationship between IDS and music. In other words, the putative ancestral ID

vocalizations did not necessarily evolve into music (at least not exclusively) through the mother-infant interactions. The contemporary IDS sounds "literally [...] musical" to Dissanayake (2000: 394) because human mothers have fully developed musical capacities. However, there is no reason to assume that the origins of modern musicality are to be found in IDS just because hominins realized that IDS is "emotionally affecting and functionally effective" (Dissanayake 2000: 401). As is the case with Falk (2004), Dissanayake is also unclear about how "the biologically endowed sensitivities and competencies of mother-infant interaction" (Dissanayake 2000: 401) actually spread to the other members of the community. Traits that are innate in infants are actually innate in humans, and "it is not necessarily the case that these innate abilities were selected for because they only directly benefit infants" (Morley 2013: 212). Also, emotive speech is commonplace in adults, and it is characterized by variations in prosodic patterns, which help to establish rapport in social interactions and achieve effective pragmatic communication (see Acosta and Ward 2011). Therefore, adjusting the emotional nuances through the prosodic properties of utterances may be selectively useful after one goes through childhood (Morley 2013: 212; see also Acosta & Ward 2011: 1146).

Fitch (2013: 497) builds on the arguments of Falk and Dissanayake and extends kin communication more broadly to "adults and their younger kin", stressing that kin selection "neatly explains the early ontogenetic appearance of language in infants". In an effort to explain the linguistic and musical capacity in both sexes, he mentions the role of fathers, who could support the mothers in parental care and complement the ID vocalizations (Fitch 2010: 493). Like Dissanayake (2000: 390), Fitch (2010: 493) also points out that parent-infant vocalizations among hominins were probably interactive; the infants were not just passive listeners. He further notes that semantics could have emerged in this context, as children today have a natural proclivity to attach meaning to the utterances spoken by their parents (Fitch 2010: 494). However, he falls in the same trap as Falk in trying to explain the phylogenetic development of music and language through the ontogenetic behavior of modern human infants. Children today find meaning in their parents' speech because they have evolved the full capacity for language and because there is a meaning to interpret. Ontogenetic processes cannot explain the selective pressures that have been at play at the phylogenetic level. Despite

his emphasis on parent-offspring vocalizations, Fitch (2010: 504) does not overlook the possibility that protolanguage could have also evolved through adult-to-adult interactions as "a manipulative, emotionally grounded vocal communication system".

The proposals reviewed in this section provide support to the view that a prosodic protolanguage was the precursor of modern language and music. By directing our attention to modern IDS, they give a hint of how early hominin vocalizations might have sounded like. Importantly, they are backed by a large body of developmental evidence that makes clear the significance of prosody in modern human communication and, possibly, in the communication of our distant ancestors (see also section 2.3). They stress the significance of intonation and rhythm in communication with prelinguistic infants, indicating that the prosodic features of language are evolutionarily more ancient than the syntactic ones. Hence, their contribution to solving the puzzle of the origins and nature of language cannot be denied.

On the other hand, there are no principled reasons to believe that early human interactions should have been restricted to mothers-infants or parents-offspring (see also Morley 2013: 209-212). If small-scale societies of indigenous populations can tell us anything about earlier hominin communities, then we should not overlook the fact that virtually "everyone in the community *is* kin" (Dunbar 2016: 274; see also Fitch 2006: 202). Additionally, nonhuman primate communities are organized in "a highly structured network of individuals linked to each other through ties of kinship, friendship and obligation" (Dunbar 2016: 24). Therefore, it is probably fair to conclude that early hominin interactions included the rest of the community members (and not just parents and their progeny) and involved a certain amount of reciprocity. Nonetheless, it is possible, as Falk (2004) suggests, that the mother-infant interactions early in the *Homo* lineage may have initially been under selective pressures that were later generalized to the adult community. As for the role of paternal vocalizations that Fitch (2010: 493) refers to, it strikes me as unlikely that they were of much significance prior to the appearance of anatomically modern humans, i.e., *Homo sapiens*. Dunbar (2016: 339-

341) provides evidence that archaic humans[17] (i.e., *Homo heidelbergensis* and other *Homo* species that preceded the emergence of *Homo sapiens* [see Fig. 3]) were polygamous, and that reciprocated pair-bonding became manifest only in *Homo sapiens*. Hence, it seems that kin selection *sensu lato* has difficulty in accounting for the phylogenetic origins of music, language and meaning. It would be more parsimonious to suggest that kin selection involving male parental care became more relevant with the advent of anatomically modern humans and the evolution of social monogamy. Thus, I see kin selection theories more fitting as part of the broader social brain hypothesis discussed below.

3.1.3. Social selection

Social selection considers an individual's ability to form close bonds with other community members. Those who are better at managing social relationships (including sexual relationships) will attain high social status, which may confer them high survivability and reproductive success (Miller 2001: 445). Hence, social selection will favor individuals who are adept at forging and maintaining long-term relationships of trust and reciprocity (Seyfarth & Cheney 2017: 82). Those who manage to be respected and trusted will receive favorable treatment and will also be more desirable to members of the opposite sex.

Group selection, on the other hand, may occur in small, cohesive groups, where the members' altruistic behavior may benefit the whole social unit (Rogers 2016). Even though a trait elaborated by group selection does not immediately benefit the individual, it improves the overall fitness of the group, enabling its members to perpetuate their genes. Because group selection seems to occur in tight and cohesive groups, it may be regarded as a loose interpretation of kin selection. Therefore, social selection as described here is not to be equated with group selection; it is not about altruistic behavior (which, from a neo-Darwinian perspective, should only be directed towards kin), but about social commitment, reciprocal bonding, or a tit-for-tat exchange of information.

[17] There is no consensus on which species should be designated under the term *archaic humans*. By *archaic humans* I mean hominins that evolved after *Homo ergaster* and that preceded the emergence of *Homo sapiens*. Thus, archaic humans lived roughly 600,000 - 200,000 years ago (see also Fig. 3).

Theories of evolution that emphasize the social aspect of language and music may vary in their interpretation of the putative selective pressures (enhanced communication, cooperation, group identity, group cohesion, social bonding, or complex social structure), and it is not always clear whether they make any distinction between social and group selection (e.g., Mithen 2005). This of course does not cancel the validity of these two processes of natural selection, as they could have both influenced the evolution of language.

Richman (1993) is an early proponent of musical protolanguage, positing that it evolved through social communication. In a brief article (1993: 722), he argues that singing lies "halfway between expressive sounds and speech". He comments that intonation, melody and rhythm are shared by singing and speech. Richman (1993: 721-722) states that singing gives a "feeling of social cohesion and solidarity" and expresses emotions more efficiently than speech, which makes it a good candidate for an intermediate evolutionary stage between affective, "primate-like vocalizations" and modern human speech. He builds on his work with the gelada baboons (Old World monkeys), whose "songlike vocalizations [...] express emotions, enhance group solidarity, and accomplish social ends" (Richman 1993: 722). Richman's model places emphasis on the musicality of protolanguage, drawing inspiration from the intricate vocalizations of geladas.

Dunbar (2016: 209) affirms that geladas have a large repertoire of complex vocalizations, unlike any other nonhuman primate, and he points out that their calls occur both synchronously in choruses and also antiphonally in "call-and-reply sequences between grooming partners". Dunbar (2016: 208-209) also notes that, compared to other primates, the social group size of geladas is "unusually large". Therefore, their elaborate musical vocal interactions function as a kind of "vocal grooming" that allows community members to keep contact through distance (Dunbar 2016: 209) and preserve social cohesion (Richman 1993). The complexity and musicality of the gelada vocalizations, along with the surprisingly large social groups, could provide some hints regarding the origins of a musical protolanguage in the hominin lineage (see Dunbar 2016: 209-210). Richman and Dunbar highlight the above social functions of the gelada vocalizations and use them as a model for the emergence of human communication. The challenge in likening nonhuman primate vocal exchanges to human speech is that

the gelada calls are innate (no nonhuman primates are known to be vocal learners). This is probably why Richman and Dunbar envisage early hominins as singing creatures before evolving to speaking creatures; an enhanced capacity for voice modulation may have preceded the ability to imitate acoustic signals from conspecifics (see also Pisanski et al. 306-308).

Mithen (2005) also emphasizes the importance of musicality in the evolution of language, believing that musical behaviors emerged early in hominin evolution. He further discusses the role of prosody (Mithen 2005: 24-25, 62), which is an important link between music and language. He imagines musical protolanguage as "[h]olistic, manipulative, multi-modal, musical and", at a later stage, "mimetic" – forming the acronym "Hmmmmm" – (Mithen 2005: 172). Moreover, Mithen (2005: 87-94) stresses the role of emotions (most explicitly expressed through music), arguing cogently that they are crucial in rationality, decision-making, social success, and critical thought. Thus, according to Mithen (2005: 179), musicality does have survival value "as a means of communicating emotions, intentions and information". He sees "selective pressures for enhanced communication" as the main driver behind hominin brain expansion and, consequently, the emergence of music and language (Mithen 2005: 158). He concludes that referential, analytic language evolved with the emergence of *Homo sapiens*, leaving the emotive aspect and "the forging of group identities" to "Hmmmmm", which in turn evolved into music (Mithen 2005: 266-267).

As it has been noted by other commentators (Botha 2009: 73; Morley 2006: 101-102), an issue with Mithen's proposal is terminology. Mithen (2005: 11, 100) is rather vague on what can be qualified as 'music', and his distinction between music and 'Hmmmmm' remains unclear throughout the book. He points out that music-making, which has principally been a shared activity in most human cultures, has many beneficial effects, such as sexual advertisement to prospective partners, support in children's "cognitive and emotional development", social bonding and a sense of "boundary loss" (Mithen 2005: 205, 208-209). He is not clear though whether earlier *Homo* species made music and danced together (in a similar way that tribal people nowadays do), or they simply communicated via a prosodic protolanguage/protomusic. This makes his argumentation somewhat circular, since the nature of 'Hmmmmm' is backed by

observations on music, and at the same time music is explained as a derivative of 'Hmmmmm' (Morley 2006: 101). This confusion has probably led to misinterpretations regarding the nature of musical protolanguage (cf. Fitch 2013: 498-499) and uncertainty as to whether Mithen (2005: 69, 100, 245, 266-267) considers that music preceded language or vice-versa.

Mithen's work deals with a range of possible selective pressures that might have spurred music and language evolution. Mithen (2005: 87-88) considers communication pressures pivotal and he links the need to communicate to social group size. Furthermore, he grounds the origins of music and language in human intelligence and brain size (Mithen 2005: 127-128). He also recognizes the significance of climatic changes and the impact they could have had on food availability, dietary shifts, bipedal locomotion, and, by extension, vocal control and brain enlargement (Mithen 2005: 145-146, 150-153). Despite his emphasis on social factors, he also considers sexual selective pressures (Mithen 2005: 176-191) and mother-infant interactions (2005: 196-204), which may have influenced the evolutionary trajectory of hominins. Thus, Mithen intertwines a wide array of evidence that provide support to the musical protolanguage hypothesis. However, in search of support for 'Hmmmmm', he is not always critical of others' ideas, and he occasionally drifts to pure speculations about the lifestyle of early humans and their putative protolanguage (e.g., Mithen 2005: 195, 204). Mithen's proposal would have probably been more persuasive had he elaborated in further detail on the selective pressures that he believes were crucial to the development of music and language, namely, what exactly is meant by "enhanced communication" and how it influenced the evolution of language and music.

Mithen adopts a multidisciplinary perspective, which I believe is necessary in order to come closer to solving this extremely complex and multifaceted problem. Although he (2005: 6) recognizes that convergent evolution may explain the capacity for complex vocal learning in multiple species, including humans, he focalizes on the evolutionary history of the hominin lineage. Mithen tries to address the different phases of human evolution, and he seeks to identify the possible selective pressures that may have been at play. Importantly, he brings to the foreground the widely overlooked role of emotions in making the right decisions, dealing with social affairs, and behaving rationally. For example, when two people love each other, they can make far-reaching

commitments, such as starting a family and sharing property, without fearing that they will be abandoned "as soon as someone physically more attractive or wealthy comes along" (Mithen 2005: 88). The observation that emotions are more deeply involved in rational thinking and intention-sharing than generally assumed urges us to appraise their adaptive value and to reconsider their traditional association with music and dissociation with language. Mithen's evolutionary scenario can be complemented by other theorists who have elaborated more on the nature of specific selective pressures.

Dunbar (2017: 211) concurs with Mithen, hypothesizing that language evolution went through a musical stage and that fully-fledged language evolved rather recently. The crux of Dunbar's (2017: 209) argument is that social group size is a principal mover in the evolution of human cognition. Laughter, singing, and language supposedly evolved in this order chronologically as a form of *social grooming* in increasingly larger hominin social groups (Dunbar 2016, 2017). Dunbar (2016: 167-170; 2017: 210) contends that laughter evolved very early, probably with the appearance of early *Homo*, because it is spontaneous, instinctive, contagious, and fundamentally social; it releases endorphin, it does not require language, and it is shared with chimpanzees. Provine (2017) adds that bipedalism was critical for the evolution of speech. Bipedal locomotion allowed for better coordination of breathing, which in turn enabled hominins to evolve "human-type laughter", complex vocalizations, and eventually speech (Provine 2017: 241-242). According to Dunbar (2017: 210), laughter in the form of "amusical chorusing" would have sufficed as a social grooming/bonding activity in the relatively small community size of early *Homo*.

With the advent of archaic humans and their more complex social structures, laughter as a form of social grooming would not have been adequate (Dunbar 2017: 210). Thus, as stated by Dunbar (2017: 210), "musical chorusing" in "a form of wordless singing, or humming" took the role of bonding group members. This may have later evolved into music-making, including rhythm and dancing (Dunbar 2016: 212). Dunbar makes clear the significance of vocal control in the emergence of speech, tracing anatomical changes that allowed breath control and enhanced articulatory capacity (Dunbar 2017: 210). He neglects to address the need to evolve learned vocalizations though; the ability for vocal learning, in addition to enhanced vocal control, is a

prerequisite for the emergence of human speech. Like laughing (and grooming among primates), singing or "doing music in any form" induces an endorphin rush, which may be augmented by synchronous movements (Dunbar 2016: 210-211). Contrary to laughter though, music appears to have a bonding effect on an unlimited number of people (Dunbar 2017: 210).

As for the emergence of language, Dunbar (2017: 211) postulates that the regular use of fire around 400,000 years ago was a key factor, as it reduced the pressure on time for bonding during the day. Thus, hominins could have focused on foraging and other necessary activities such as tool-making during the day, and leave socializing for the evening hours around a hearth (Dunbar 2017: 211). This drove Dunbar (2016: 232-233) to speculate that the poor visibility around a campfire would have prompted vocal exchanges, exerting selective pressures on language, which "would have allowed considerable information exchange" about social relations in the dim light. Although I doubt that fire alone was the triggering factor for the appearance of language, it could have facilitated communication after the sunset, and it thus may have influenced the trajectory of language evolution. Moreover, Dunbar (2016: 270-271) asserts that story-telling was a prime driver for language, mainly because of its effect on bonding the whole community and creating a sense of identity. Stories are probably an ancestral characteristic of language, but they seem to be the result of having language rather than spurring the evolution of language. Additionally, in favor of "everyday social uses of language for social exchanges and story-telling", Dunbar (2016: 234) dismisses "the symbolic use of language" as of having evolutionary importance, claiming that symbolic language appeared much later, possibly 50,000 years ago. However, he overlooks the fact that story-telling (as he describes it) requires symbolism. His view that exchange of social information influenced language evolution has probably more explanatory power than story-telling, and it seems evolutionarily plausible. Human language is by definition symbolic. A critical point is how and when symbolic language emerged out of more primitive forms of communication (discussed in section 3.3).

The backbone of Dunbar's argumentation is the *social brain hypothesis*. It predicts that behavioral and social complexity in primates correlates with brain size,

especially the neocortex[18] (Dunbar 2016: 59-64). Observation of nonhuman primate behavior demonstrates that there is a relationship between neocortex volume and behavioral complexity, such as "grooming clique size, the use of coalitions, [...] deception, [...] mating strategies [...] and the complexity of facial and vocal repertoires" (Dunbar 2016: 61-64). Thus, the social brain hypothesis provides a compelling model for predicting social complexity in the various hominin species based on brain size (Dunbar 2016: 66). The social brain hypothesis allows us to estimate behavioral patterns in hominin evolution and to make principled predictions about when more complex forms of communication, such as language, might have evolved. It is not clear though why the emergence of language became evolutionarily necessary, since, according to Dunbar (2017: 210), singing has no limit on the group size it can bond, and "language evolved to create and bond small, exclusive communities" (Dunbar 2016: 272). In this line of thinking, language should have evolved first, followed by music. These cast doubts on Dunbar's conviction that singing preceded language and question the validity of his argumentation regarding the selective pressures (e.g., fire) that induced the emergence of language. Hence, it would probably be more parsimonious to proffer that language and music/singing coevolved in parallel in increasingly larger social groups (De Boer 2017: 160; Morley 2013: 225; Brown 2017: 4).

Nevertheless, Dunbar's (2016: 271-277) arguments illustrate the importance of kin and cohesive groups of related people for vocal communication and social information exchange. He discusses the nature of kinship and argues convincingly that the boundaries between kin and nonkin are fuzzy in small, traditional communities (Dunbar 2016: 272-277, 316-317). Thus, he merges neatly the relevance of kin and group selective pressures. In addition, his advocacy for social selection does not mean that individuals evolve a trait at the expense of themselves and the benefit of the group (Dunbar 2016: 312). Rather, Dunbar (2016: 39) sees social grooming in primates as a way to defuse "the stresses of group-living". Therefore, social bonding should not necessarily be associated with altruistic behavior. Indeed, pair-bonding – which could be interpreted as an important step towards the social complexity that characterizes

[18] The neocortex is a thick, layered structure of the outer part of the mammalian brain, and it is involved in higher-order cognitive functions.

humans – could be understood as a socially selected behavior that helps an individual obtain the desired status and improve her/his fitness by building and maintaining long-term bonds of trust and reciprocity. It may have evolved to counter infanticide risk (Dunbar 2016: 332), which mitigated the danger of harassment for females and allowed "males to monopolize matings with at least one female" (Dunbar 2016: 335).

Oesch (2019) emphasizes the adaptive significance of music and language in the context of social interaction. He draws on Dunbar's social brain hypothesis and stresses that human language and music do "not normally occur in a social vacuum" (Oesch 2019: 3). He attempts to use convergent evolution in support of the social functions of music, citing studies on the vocalizations of some bird species and primates, notably gibbons (Oesch 2019: 3-4; cf. section 2.4.1). Even though among birds these vocalizations occur chiefly during courtship, in many mammals, including primates, they have social functions (see also Filippi 2016). Furthermore, the antiphonal vocalizations of various mammals (Oesch 2019: 4-5) indicate that the sense of turn-taking might be an ancient trait that evolved long before the rise of primates. I agree with Oesch (2019: 5) that vocal turn-taking in humans and other unrelated mammals could be seen as "an instance of convergent evolution", where the prime function is possibly to handle social relationships.

It has been argued that singing, along with coordinated movements such as group drumming or dancing, promotes social bonding (Dunbar 2017: 210; Oesch 2019: 4). Experimental research on joint singing and dancing among young children confirms music's bonding effects (Kirschner & Tomasello 2010). These effects could be attributed to features specific in music, such as entrainment to a periodic pulse, or "the use of *discrete pitches*" and a "*repetitive melodic structure*" (Kirschner & Tomasello 2010: 357). Kirschner and Tomasello (2010: 362) argue that music encourages cooperation and creates a sense of unity without being oriented towards a goal. That is, they understand music as a "*biocultural phenomenon*" – rather than the outcome of pure biological evolution – whose components originally evolved for other adaptations (Kirschner & Tomasello 2010: 362). Nonetheless, as it has been discussed above (see Mithen 2005: 87-89; Acosta & Ward 2011), the expression of emotions (whether communicated through music or other means) can have an adaptive value. Furthermore, Oesch

(2019:5) points out that even in human conversations where turn-taking is the norm, synchrony is obvious in several aspects, including lexical choice, grammar, phonology, and speech rate (see also Levinson 2016). Oesch (2019: 5) interprets this effort to improve the success of a conversation as a desire to "increase affiliation with a conversation partner". In short, social bonding in both language and music helps to establish rapport and deal with social encounters.

Finally, Redhead and Dunbar (2013) adduce experimental evidence arguing that language originally evolved as a social tool, namely, to manage social relationships. They argue that it is implausible that language evolved to facilitate cooperation and help transmit information for ecological problems such as tool-making or food location; many animal species manage to cooperate efficiently and inform conspecifics about predators or food sources without language (Redhead & Dunbar 2013: 845-846). In their recall paradigm experiment,[19] the participants remembered social information better than factual content or the flamboyant language used in sexual advertising (Redhead & Dunbar 2013: 849). The results led them to conclude that social information exchange is essential to language function, whereas sexual advertising or "instrumental uses of language" may be emergent properties (Redhead & Dunbar 2013: 850). Although this experiment only provides some hints regarding the primary functions of language, social communication seems to be central. Sexual advertising or flamboyant speech might simply be cultural artefacts instead of products of biological evolution.

After the split from the LCA with the chimpanzees, hominins kept evolving for several million years. Therefore, it is likely that different selective pressures influenced the trajectory of human evolution. The selective pressures discussed so far need not be mutually exclusive; all of them could have played a role at some point or another. The evidence suggests that sexual selection did not have a role to play. If it did (possibly in the early stages of hominin evolution), it was a minor one. Also, kin selection cannot be effectively dissociated from group selection; despite their differences, they both

[19] The participants were asked to read a short paragraph and then jot down as much information as they could remember (Redhead & Dunbar 2013).

describe altruistic behavior towards related individuals. Kin/group selection may not be the sole answer, however. As the bonds of genetic relatedness weaken in increasingly larger social units, socially selected traits may make their appearance; individuals will try to improve their fitness (and the fitness of their close relatives) by manipulating others' behavior and by creating alliances. Hence, even if kin selection triggered the evolution of articulated speech, the increasing size of hominin groups would have influenced its evolutionary path. Almost all extant primates have "complex and highly structured social systems", and we can thus be confident that sociality was a long-established feature of early hominins (Fitch 2010: 228). Consequently, the social selective forces can hardly be ignored.

3.2. The "musilanguage" model

Brown (2000, 2017), an advocate for a common precursor of music and language, emphasizes the importance of prosody and emotion in the evolution of human speech. He conjectures that a prosodic protolanguage, which he terms "musilanguage", was the means of communication earlier in human evolution, initially in the form of "ritualized territorial chorus of emotional communication" (Brown 2017: 5). Although he does not indicate at which point in human evolution musilanguage could have emerged, he proffers that it was neither musical nor linguistic, but rather it included distinct ancestral features that are common to music and language (Brown 2017: 277-278). This communication system, which became increasingly dependent on prosody, eventually bifurcated into music and language, the former being characterized by tonality and the latter by lexicality (Brown 2017: 10).

Brown (2017: 2-5) outlines two stages of musilanguage. The first is characterized by "affective prosody": innate calls which are "phylogenetically derived from primate communication" and hence are similar to a nonhuman form of group chorusing (Brown 2017: 8). The second phase is comprised of "intonational prosody", which is "a system of nonsense vocalizing or pure prosody" (Brown 2017: 6). During this second phase, humans evolved in parallel the capacity for vocal control and vocal learning, as well as enhanced affective prosody that communicated more accurately the intensity of emotions (Brown 2017: 6-8). Furthermore, Brown (2017: 7-8) writes that "phonemic

combinatoriality" was a key feature in this musilanguage phase, but since lexicality appeared only with the emergence of fully fledged language, the learned vocalizations were "devoid of intrinsic meaning". Brown (2017: 8) also claims that musilanguage enabled communication via holistic intonational phrases, but these were strictly "prosodic holophrases" and not symbolic ones (see also section 3.3). Nevertheless, Brown (2017: 9) asserts, "the same phrases would be uttered repeatedly", mainly "during group chorusing".

The first stage of *affective prosody*, although not analyzed in much detail, seems evolutionary plausible as the logical continuation from nonhuman primate vocalizations. By modulating voice pitch and timbre, our ancestors presumably conveyed basic emotions and intentions. However, the *intonational prosody* stage contains some inconsistencies. One may justifiably wonder how there can be phonemes without any meaning, since they distinguish one lexical item from another. Also, if this was a system of "pure prosody than pure phonology" (Brown 2017: 8) then it is not clear why there were selective pressures for learned vocalizations and why the same vocalizations were performed repeatedly in groups. Indeed, Brown gives no explanation for this. He (2017: 8) briefly likens the purported meaningless vocalizations to birdsong. However, birdsong is profoundly different from human language and music; as discussed in section 3.1.1, birdsong's main function is sexual advertising, and it has not served as the precursor for a cognitively more demanding communication system. Additionally, Brown (2000: 296-297) mentions that music evolved through group coordination and cooperation, but he doubts that a neo-Darwinian explanation can account for music's emergence, pointing to cultural group selection. Thus, he does not address any selective mechanisms in the biological evolution of music and language.

Furthermore, Brown (2017: 14) states that the "prosodic holophrases" he argues for scaffolded the emergence of "compositional syntactic mechanisms". He correctly emphasizes the significance of prosody in human communication and the possibility that it had a major evolutionary role, preceding syntax. However, prosody alone is insufficient in explaining how analytic language evolved. The "prosodic scaffold" that Brown (2017: 13-14) proposes does not provide any evolutionary explanation and is vague on how syntax could have possibly appeared. In addition, there seems to be a lack of a gradualist perspective in the bifurcation of musilanguage to music and language.

Purely linguistic features such as propositional meaning and lexical items probably emerged relatively late in human evolution. Since by *language* Brown (2017: 9) means "full-fledged language", it is hard to imagine a leap from solely prosodic holophrases to language (without any intermediate stage of symbolic holophrases consisting of phonemic components).

The complete absence of a time frame makes difficult to assess the plausibility of Brown's complex theory. Moreover, the confusing terminology, the vagueness about the process of bifurcation, and the incongruity in some of Brown's ideas add to the problem (see also Botha 2009: 69-71). Finally, I think that it is not necessary to decide whether some properties are strictly musical or linguistic (see section 2.1; cf. Brown 2000: 277-278); rather, we should focus on explaining why a particular property emerged and what function(s) it serves, whether employed in music or language (see also Fitch 2010: 488). In conclusion, despite the various obscure parts in Brown's model, he succeeds in drawing attention to the evolutionary importance of prosody in human communication. He rightly points out that most of the contemporary languages are tonal (Brown 2017: 11; see also Patel 2008: 40; Fenk-Oczlon 2017: 2; section 2.1.1), and he lists a number of studies which make clear that prosody conveys more than just the expression of emotions.

3.3. Holistic protolanguage

The premise of several musical protolanguage hypotheses is that protolanguage was holistic in nature (Jespersen 1922; Mithen 2005; Fitch 2010; Brown 2017). The idea dates back to Jespersen (1922: 439-441), who argues for a gradual evolution of language. Jespersen (1922: 421-422) notes that languages tend to become more analytic, evolving from indivisible clusters to shorter components that can be combined. He thus proposes that language originally consisted of "indissoluble rigmaroles, with [...] a dim meaning attached to them" gradually evolving "word-like smaller units [...] capable of being analyzed and combined with others" (Jespersen 1922: 439-441). In support of his argument, he points to the way children acquire language: they try to identify individual constituents (i.e., words) in adult utterances, and associate particular sounds with meaning (Jespersen 1922: 439-441). Moreover, Jespersen (1922: 419-421) draws

attention to the ubiquity of tonal languages, which were probably even more common in the past, since several languages have lost their "pitch accent". He also mentions that the songs of many indigenous people "contain totally meaningless syllables" (although this is not restricted to indigenous populations), and that "rhythmic singing" is a universal practice among the world's cultures (Jespersen 1922: 434-437). Therefore, Jespersen (1922: 441-442) concludes, language "began with half-musical unanalyzed expressions".

Jespersen refined the musical protolanguage hypothesis originally mooted by Darwin. The suggested progression from inseparable clusters to analytic language adds to the plausibility of musical protolanguage. If the ontogenetic origins can tell us anything about the phylogenetic origins of language, the developmental evidence from child language acquisition (e.g., Tomasello 2003: 36-40) confirms Jespersen's view. Although Jespersen does not explain in detail how this transition could have happened, his effort is an important step in explaining how meaning originated (see also Fitch 2010: 484). Additionally, his observations based on comparative data (especially on autochthonous cultures) are to the point, offering support to the idea that protolanguage was musical in nature. Also, the fact that many languages were (and still are) tonal remains an issue worth considering, making clear the importance of prosody. Even though they did not prompt further discussion for several decades, Jespersen's insights regarding a protolanguage that was both prosodic and holistic are still felicitous one century later.

A contemporary fervent proponent of a holistic protolanguage is Wray (1998, 2000). She observes that formulaic language is ubiquitous, typically used in "common socio-interactional exchanges", whereas analytic language is employed in thinking and in developing new ideas (Wray 2000: 300). Her conviction is strengthened by the fact that chimpanzees use manipulative holistic vocalizations and gestures in requests, greetings, threats, or offers (Wray 2000: 289), and also by the fact that during L1 acquisition children learn formulaic sequences before grammar (Wray 1998: 61). Furthermore, she argues that our "short-term memory capacity" seems "to be insufficient for our powerful analytic faculties" (Wray 2000: 299). Hence, she concludes that grammar did not exert evolutionary pressures in early human speech. Instead, "word-sized concepts"

and syntactic language arose as a way "of organizing creative thought and planning" (Wray 2000: 291). Protolanguage was not descriptive or referential because, according to Wray (1998: 51), there was no need for it; protolanguage vocalizations contained broad and functional meanings with no internal structure, which were used in daily social communication.

Wray (2000: 296) assumes that protolanguage evolved into synthetic language through the process of "segmentation". She believes that the holistic sequences carried latent "meaningful subunits" that could be extracted, splitting the holistic phrases into words (Wray 2000: 296-297). Moreover, Wray (1998: 55-58, 2000: 297-298) gives concrete examples of how holophrases could have segmented, purporting that chance alone would have prompted segmentation in cases of phonetic similarity between constituents of different holophrases. To support her claim, she compares this process with language acquisition; insights into the way languages are acquired show that holistic processing may be pertinent to language evolution. However, the argument becomes invalid when it is extended in order to explain the transition to compositional language; children are able to extract meaningful components from longer sequences because there are separate meanings in those sequences. It is highly doubtful that holistic protolanguage sequences contained "meaningful subunits" that that were later segmented into words. It probably makes more sense to suggest that languages were reinvented and gradually transformed from their protolanguage predecessors. As humans acquired the ability for referential and descriptive expressions, they began to incorporate lexical words in their protolanguage, instead of finding patterns in preexisting holistic phrases. This explanation would allow for a more gradual transition since it would not have rendered speakers dependent on the cultural evolution of protolanguage. Thus, in the transitional phase, analytic language was evolving alongside the waning holistic protolanguage. Wray (1998: 57) acknowledges this possibility, arguing that after "a critical mass point" analytic language dominated over the holistic protolanguage. Nonetheless, holistic processing has not disappeared completely; formulaic language is much more pervasive than one might initially think, perhaps being a remnant of our evolutionary history (Wray 1998: 64-65; Mithen 2005: 277).

Although Wray does not discuss the possibility of a protolanguage based on prosody, Mithen (2005) incorporates her model into his musical protolanguage. He

states that the expression of emotions was a primary characteristic of protolanguage; the communication system of our distant ancestors was not based only on holistic utterances, but also on musical qualities like rhythm, pitch, and timbre (Mithen 2005: 172). Additionally, the existence of a holistic system for a prolonged period could explain the stability and success of *Homo ergaster*/*erectus* (Fig. 3) in surviving for more than a million years, and at the same time it could explain their cultural stagnation and lack of innovation (Mithen 2005: 173). The appearance of compositional language had an enormous impact on human consciousness and creative thinking, setting off imagination and cultural innovation (Mithen 2005: 263-265; see also Wray 2000: 290-291).

Mithen (2005: 255-257) and Fitch (2010: 501-503) cite Kirby's experimental work as strong evidence in favor of a transition from a holistic protolanguage to analytic language (see Kirby 2000, 2001). Mithen and Fitch mention that in Kirby's computer simulation models and iterated learning experiments a form of compositional language evolves out of an input which consists of structureless and meaningless utterances or strings of symbols. This transformation of such experimental mini languages occurs within a few generations of learners/agents. Mithen (2005: 255) claims that Kirby's "simulations are able to explore [...] how children acquire language", even though no children participated in the cited experiments. Moreover, both Mithen and Fitch seem to overestimate the pertinence of cultural evolution in the quest for the biological origins of language. Kirby's works does not "effectively simulat[e] the process that [...] may have happened during human evolution" as Mithen (2005: 256) maintains. Fitch (2010: 502) is more cautious, acknowledging that Kirby's simulation models predict cultural and not biological evolution. Nevertheless, he argues that cultural change would influence biological evolution, favoring those who could "master the new analytic system more rapidly" (Fitch 2010: 502).

Kirby's work demonstrates that the transition from a holistic to a compositional system of communication is possible, but one should not hasten to overgeneralize these findings. If one applies cultural change models to the biological evolution of language, they presuppose that the presumed holistic protolanguage speakers had the same cognitive abilities and proclivity for synthetic language as modern humans do. In addition, the transition from protolanguage to language would have certainly required

many more generations than just the handful needed in the iterated learning experiments. Therefore, it is very possible that biological selection would have had a say, unlike Kirby's models where there is not any natural selection (cf. Fitch 2010: 501-502). Finally, if one accepts the possibility that a prosodic, holistic protolanguage lasted for an evolutionarily significant period of time, it is sensible to posit that there were different phases that characterized it. Thus, a strictly prosodic protolanguage with a strong affective component and no vocal learning (cf. Brown 2017: 8-9) could have been succeeded by a holistic protolanguage which consisted of a "culturally transmitted vocal repertoire" (Fitch 2010: 497) with even more nuanced emotive and pragmatic meanings. On the way to fully syntactic language, there was probably a period where holistic, prosodic phrases were gradually giving way to "word-sized concepts" (Wray 2000: 291), referential function, and symbolism.

Addressing the selective pressures that drove language evolution is a complex and interdisciplinary endeavor. The biological adaptations that underpin the capacity for language evolved over many thousands (or millions) of years. Therefore, it is likely that the evolution of language has been shaped by different selective pressures. Theories that focus on *sexual selection* are insufficient to explain critical aspects of language. Nevertheless, through the instances of convergent evolution, they underscore the role of vocal imitation for language to be learned and transmitted (as Darwin first did). Moreover, the analogy with bird or whale song shows the importance of *vocal modulation* (manipulation of pitch and timbre) in communication with conspecifics.

Kin selection accounts remind us of the musical nature of IDS. Ample developmental data illustrate the pivotal role of ID speech and singing in interactions with infants; prelinguistic infants are very sensitive to the prosodic features of ID vocalizations, which scaffold their language acquisition. Infants tune in quickly to the prosodic properties of their language, and they start babbling in its prosodic patterns before they start speaking. Additionally, kin selection fits well with the fact that humans are born altricial and go through a prolonged childhood. Their extended period of helplessness may have selected for an ability to imitate maternal vocalizations that are characterized by distinct musical qualities (wide range of pitch variation, intonation contours, rhythmical patterns) and a strong affective component.

Proposals that explain language as a socially selected trait also point to the need for voice modulation and enhanced vocal control. Individuals who were able to produce novel sounds and manipulate pitch contour, timbre, and intensity in their vocalizations were presumably more skillful at managing their social relationships. As this would have improved their fitness, those who were competent at imitating them would have also improved their survivability and reproductive success. Language evolution through *social selection* and the *social brain hypothesis* imply that increasing social complexity required more than affective vocalizations regulated for pitch, timbre, duration, or intensity; at some point, the need to convey referential and propositional meaning ushered the development of syntactic complexity.

The various *musical protolanguage* proposal discussed here rely on voice modulation, vocal control, and vocal learning to explain the origins of language. The ability to volitionally modulate one's vocalizations presumably enabled hominins to communicate affective and pragmatic meaning more efficiently. The idea that modern language evolved from a musical/prosodic precursor is supported by diverse findings (see section 2) and is consistent with the notion of a protolanguage that was *holistic* in nature: since it is argued that protolanguage originally conveyed only emotive and pragmatic nuances, there was no need to have lexical items and syntactic structures. The prosodic modulation of the voice and the utterance of "half-musical unanalyzed expressions" (Jespersen 1922: 441-442) must have sufficed in the early stages. A corollary of a protolanguage that was both musical and holistic in nature is that grammar emerged rather late, roughly coinciding with the appearance of modern language (see Tomasello 2003: 13-14).

4. Objections to musical protolanguage

A harsh critic of the musical protolanguage hypothesis is Tallerman (2006, 2013), who has attacked virtually every aspect of it, especially the holistic nature (see section 3.3). Tallerman (2013: 456) disagrees with Fitch (2010) and alleges that one cannot learn anything about language evolution by studying other animals. While it is true that Fitch (2010: 471-472; 2013: 494-495) might be placing excessive importance on analogy (especially on learned vocalizations in birds and other animals unrelated to humans),

one should not simply dismiss the comparative approach. The study of hominid[20] communication and cognition is of greater significance than Tallerman would have us believe. Chimpanzees, with whom we share our last common ancestor (LCA), are of particular interest. Understanding how primates that are closely related to humans behave and communicate with each other may yield insights in the evolution of communication strategies in the hominin lineage. Tallerman (2013: 457) apparently criticizes Fitch for comparing species instead of comparing languages "in order to deduce properties of a common ancestor". But there is a limitation to how far back in time comparing languages can take us; to get closer to the origins of language we need to consider what preceded it, and the broad comparative method is one of the most important tools we currently have (see Fitch 2017: 5-6). Making comparisons across languages can mainly shed light on the cultural and not the biological evolution of language.

Moreover, in criticizing Fitch, Tallerman (2013: 462-463) stresses the crucial role of syntax, which, as she mentions, is underrepresented in Fitch's account. However, the properties of English syntax that Tallerman discusses cannot tell us much, if anything, about the nature of a primitive form of early human communication. Additionally, Tallerman (2013: 459) emphasizes "the exceptional nature of language as a human faculty", whereas she does not seem to embrace the "multicomponent approach" (see Fitch 2017: 4) to language and the view that some "subcomponents of language can be found in the communication systems of other species" (Fitch 2010: 26; see also section 2.4; cf. Tallerman 2013: 457). Tallerman (2013: 469-470) criticizes the relevance of complex vocal learning in the evolution of language, rejecting the idea that it might have arisen before and independently of semantics. Her claim is based on computational models that suggest that in ontogenetic development a growing vocabulary leads to an expanding phonology (Tallerman 2013: 471). However, research on real humans (see Houston 2016: 51) demonstrates that young infants are able to distinguish phonemic contrasts in different languages, before they are able to acquire words from their own language. Hence, if ontogeny bears any relevance, the ability to make semantic

[20] *Hominids* are all the members of the great ape family, including humans. The term *hominins* refers to humans and extinct species of "the lineage leading to modern humans that arose after the split with the LCA" (Dunbar 2016: 8).

distinctions follows the ability to make phonemic distinctions, which in turn is anteceded by the capacity to perceive the prosodic features of language (see also section 2.3).

Tallerman is confined on the cognitive abilities of modern humans to explain the nature of the language faculty. The view that syntax is the essence of modern language does not contradict the musical protolanguage hypothesis; it is the rigid view that syntax, semantics, and phonology must have been unbreakably linked in this order which impedes any credible evolutionary explanation, pushing us to accept saltation or a sudden mutation event (see Bickerton 2000: 267). No one denies that human language is unique, but an overreliance on its uniqueness will not get us far. When one postulates the existence of a protolanguage that preceded modern language, one should be ready to acknowledge the differences the two systems must have had. Dismissing a protolanguage hypothesis because it focuses on vocal learning rather than on syntax is a groundless argument. Tallerman's approach to the biological emergence of language does not leave any space for the gradual elaboration of the various subcomponents of language through natural selection (cf. section 5).

Botha (2009) takes aim at Brown's (2000) and Mithen's (2005) protolanguage proposals, noting that his criticism is addressed specifically to these two theories and not to musical protolanguage as an evolutionary possibility (Botha 2009: 75). Although he makes some accurate observations (see sections 3.1.3 & 3.2), his argumentation is grounded on some stiff conceptualizations, which make it difficult to accept a neo-Darwinian approach to language evolution. His distinction between language ("a cognitive system") and speech ("a form of behaviour for which that system is used") is rather counter-productive in this context (Botha 2009: 65-66), as he is drawn on technicalities that miss the whole point: in order to decide on the plausibility of a common precursor of music and language we need to compare not only their neurological underpinnings, but also their acoustics and structure. If we make the reasonable assumption that the evolution of language is contingent on the evolution of speech, then it follows that any insights on the evolution of speech will have implications for the emergence of language. Botha's (2009: 66-67) adherence to rigid definitions misses the essence of the argument and does not provide constructive criticism. After all, Botha makes a language/speech distinction, but he does not discern music as a cognitive capacity and music as the acoustic manifestation of such a capacity. In this line

of thinking, and by sticking to a narrow definition of language as a starting point, one will not be able to roll back in time and assess proposals that try to explain what was happening before language and music as we know them emerged.

Jackendoff (2009) lists a number of differences between language and music, arguing that there is no persuasive evidence that the two systems evolved from a common precursor. He decides to "abstract away" from the acoustic elements that music and language share, overlooking the critical role of tonality (Jackendoff 2009: 198). Jackendoff (2009: 199) does not see any common ground in pitch between music and language, focusing instead on "the structure of phonological space" in language which he contrasts with music's "tonal pitch space". This only shows that language and music are not the same thing, but it does not investigate the possibility that they might have shared ancestral features. Additionally, Jackendoff (2009: 200) briefly mentions that "linguistic intonation and musical pitch are controlled by distinct brain areas", but he fails to acknowledge the body of evidence which suggests that prosody and musical pitch processing are channeled onto the same neural resources (Patel 2008: 72; see also section 2.2). Moreover, Jackendoff (2009: 202) admits that some structures, like hierarchical structure in music and language, could have evolved out of "a more general, evolutionarily older function", but he assumes that hierarchical structures were at some point appropriated by language and music independently. Although he recognizes possible evolutionary links between music and language, he rejects the idea of shared ancestral features. While he (2009: 203) correctly urges "caution in drawing strong connections between language and music" in the modern human brain, he emphasizes the differences in formal structure, neglecting to consider links and possible intermediate evolutionary stages in the two domains.

The holistic nature of musical protolanguage is a key aspect of several proposals (see section 3.3). Tallerman (2006, 2013) opposes vehemently the idea of a holistic protolanguage, but her argumentation mainly hinges on pointing out the differences between the postulated holistic protolanguage and modern analytic languages. Some proponents of holistic protolanguage provide some unfortunate examples of putative holistic utterances which are unnecessarily complex (e.g., Wray 2000: 294; Mithen 2005: 172) and justify Tallerman's (2006: 106) disapproval. However, infelicitous examples such as "go and hunt the hare I saw five minutes ago behind the stone at the top of the

hill" (Mithen 2005: 172) do not nullify the whole idea, as protolanguage utterances were most probably less complex. Tallerman's criticism is centered around specific examples of holistic phrases, doubting that they could have transformed into lexical items. In any case, it is rather unproductive to discuss specific examples of putative holophrases; both sides will draw on arguments that are a matter of the cultural evolution of language and have little relevance to its biological origins. One should acknowledge that any putative protolanguage (at least in its early stages) was spoken by hominins that had not yet evolved the full package of cognitive abilities that are required for modern linguistic expression. By using modern analytic language as the starting point, Tallerman cannot imagine how such a complex system could have evolved from something different. Tallerman's (2006: 107) insistence on drawing a firm line of distinction between human language and animal communication systems shows her unwillingness to adopt an evolutionary perspective.

Tallerman's arguments are rooted in Bickerton's (2000: 264) proposal of a lexical but "structureless protolanguage". In fact, Bickerton's claim of a lexical protolanguage is merely based on the preponderance of lexical items in modern languages; thus, lexical items are taken as a sine qua non for protolanguage, which originally must have consisted of a small number of "(proto)words" (Bickerton 2000: 264). Early hominins had a smaller brain size, but this does not mean that they possessed a language with the same basis as ours (i.e., words), just with a simpler or no structure, as Bickerton alleges. Tallerman (e.g., 2013: 475) focuses on pointing out the differences between modern humans and apes, neglecting to consider the cognition of extinct hominin species. The difference in brain size and consequently neuron connections between anatomically modern humans and earlier hominin species suggests that extinct hominin species communicated on a quite different level than we do. *Homo ergaster* (see Fig. 3) certainly had more limited abilities than *Homo sapiens*. Attaching symbolic meaning on specific objects or concepts requires advanced cognitive and categorization skills that are absent in nonhuman primates, and it is thus unlikely that early hominins possessed symbolic skills comparable to our own (see Tomasello 2003: 11-12). Therefore, it seems an oversimplification to relegate the communication system of earlier hominins, such as the australopithecines (see Fig. 3), to a 'dumb' lexical protolanguage in which each utterance contained only "four or five units" (cf. Bickerton 2000: 264). Bickerton (2000:

271-276) uses up space discussing English syntactic structures, and he even likens protolanguage to contemporary pidgin languages. However, this does not probe into the evolutionary origins of language. Finally, the lexical protolanguage that Bickerton and Tallerman propose seems to be against evolutionary continuity; the appearance of words and syntax is apparently attributed to catastrophic events and not to gradual elaboration of cognitive traits.

5. The emergence of language: reconstruction of its evolutionary history

5.1. Gradualism

Language is a multi-faceted and extremely complex phenomenon, and how exactly it has evolved is a matter of contention. Insights from evolutionary theory, neuroscience, language acquisition, and animal cognition and vocal behavior have helped us formulate coherent hypotheses on language evolution. Another source of important information is the fossil record: anatomical features and paleogenetic data from extinct hominins point to speech-related adaptations that have occurred throughout our evolutionary history and provide the basis for constructing a timeline for the evolution of language (see, for example, Fitch 2017: 17-24) or musicality (Morley 2013: 319-325).

Particularly among generative linguists, language has been regarded as a unique cognitive ability equated with syntax (e.g., Chomsky 2010). More recently, however, a monolithic view of language has come to be rejected (e.g., Dediu & Levinson 2013; Boeckx 2017; Fisher 2017; Fitch 2017: 4; see also Martins, Marie & Boeckx 2018: 67). Besides the mental capacity for hierarchical structuring, language also includes the ability for ostensive communication (see Scott-Phillips 2017: 187). Moreover, language is inextricably connected to speech, its physical manifestation (see also section 1). Here it is argued that language and speech evolved in tandem, involving a series of cognitive changes and anatomical adaptations in the human lineage (see de Boer 2005: 282-283, 2017: 162-163). Language, like music, is underpinned by different cognitive abilities, some of which are shared with other animals (cf. sections 2.1, 2.4). Vocal learning is a prime example (e.g., Marler 2000; Fitch 2013: 494), while the capacity to modulate one's vocal signal or entrain rhythmically may be partly shared with nonhuman animals; a

crucial difference is that humans perform these actions volitionally (cf. Clay, Archbold & Zuberbuhler 2015; Patel 2014; Pisanski et al. 2016). Thus, also here the "multicomponent approach" (Fitch 2017:4; see also Martins, Marie & Boeckx 2018: 67) is adopted, which has made it possible to set language in a broader evolutionary context.

Recognizing the plethora of mechanisms that constitute language is also necessary if one takes seriously the accepted view that evolution is gradual. On the basis of that view, the linguistic and musical abilities of modern humans must have evolved in small incremental steps over a long period on the evolutionary timescale. The cognitive skills and mechanisms involved in language will have accumulated slowly, having evolved in different periods (MacLarnon & Hewitt 1999: 342; Dediu & Levinson 2013: 10) as adaptive responses to various environmental pressures – including, possibly, climatic ones (cf. Merker, Morley & Zuidema 2015: 4). Thus, the human lineage went through several phases of preadaptations to language before reaching the current state (Dediu & Levinson 2013: 10). In fact, some of the mechanisms related to language and music are likely to have evolved before the human lineage diverged from the LCA with chimpanzees, which means that the human capacity for language was built on cognitive and physiological adaptations that are not unique to humans (Boeckx 2017: 194; Fitch 2017: 7-9; see also Morley 2013: 168).

5.2. Archeological record

The archeological record is an important source of information in reconstructing the evolutionary trajectory of the human lineage. Although physiological changes in the fossil record do not give us direct evidence about the linguistic abilities of extinct hominins, they allow us to draw inferences about enhanced cognitive abilities and anatomical adaptations for articulated speech. Hominin fossils also allow us to extract and analyze ancient DNA (Dediu & Levinson 2013: 3; Fitch 2017: 18). Recent developments in genomics[21] make it possible to trace changes that have occurred in language-related genes (e.g., Dediu & Levinson 2013: 3-5; Fitch 2017: 18-22; Martins, Mari & Boeckx 2018).

[21] *Genomics* is the field of the study of genomes. The *genome* is "the complete set of information in an organism's DNA" (Roth 2019: 443).

5.2.1. Paleontological data

Even though there is no consensus on which (if any) anatomical traits may be regarded as proof of the capacity for speech, there are a number of key adaptations that help to put the pieces of the puzzle together. It has been argued that the descent of the larynx has enabled hominins to expand the range of their vocal abilities (e.g., Morley 2013: 136). Since the larynx does not fossilize, its position can be inferred by the skull's morphology and the position of the hyoid bone (Fitch 2000: 262). The shape of the skull in fossil hominins indicates that "the laryngeal physiology of the australopithecines resemble[s] that of modern apes", whereas *Homo heidelbergensis* from around 400,000 years ago has a modern-like upper respiratory tract (Morley 2013: 140). *Homo ergaster* cranial morphology (circa 1.75 MYA) is at an intermediate stage (Morley 2013: 138). Additionally, the hyoid bone, on which "the tongue base attaches and which supports the larynx in the throat", is indistinguishable in australopithecines from that of the chimpanzees (Morley 2013: 144). The hyoid bones of modern humans, Neanderthals, and *Homo heidelbergensis* specimens from more than 500,000 years ago are virtually identical, suggesting that the process of lowering the larynx started early in the *Homo* lineage (Morley 2013: 145-146; de Boer 2017: 159). The morphology of the hyoid bone is related to the presence of air sacs, which are thought to be an impediment to speech capacity "by reducing the range of distinctive [...] sounds which can be produced" (Dediu & Levinson 2013: 6; see also de Boer 2017: 159). With the exception of humans, all great apes possess air sacs, as probably did the australopithecines (de Boer 2017: 159), but not the Neanderthals (Fitch 2000: 262). Thus, the current evidence suggests that later *Homo* species had an increased potential to produce a wider range of frequencies (and more vowel sounds), as the lowered larynx allowed for more articulatory space (Morley 2013: 142; Dunbar 2017: 210).

The hypoglossal canal (Fig. 4), which innervates the muscles of the tongue, is another source of evidence regarding the evolution of speech. The hypoglossal canal in modern humans is larger than the chimpanzees' and gorillas' canal both in relative and in absolute terms (Kay, Cartmill & Balow 1998: 5418). The canals from australopithecines and *Homo habilis* (see Fig. 3) specimens are in the range of the great apes, whereas those from later *Homo* specimens are within the range of modern humans, suggesting

that a modern size hypoglossal canal had evolved by 400,000 - 300,000 years ago (Kay, Cartmill & Balow 1998: 5418-5419). Kay, Cartmill, and Balow (1998) propose that a large hypoglossal canal supplies the tongue with a large number of motor nerves which enable it to move more flexibly. A more sophisticated control of the movement of the tongue in the oral cavity shows the increasing importance of the ability to produce complex vocalizations and may reflect the evolution of human speech (Kay, Cartmill & Balow 1998: 5417; Morley 2013: 148).

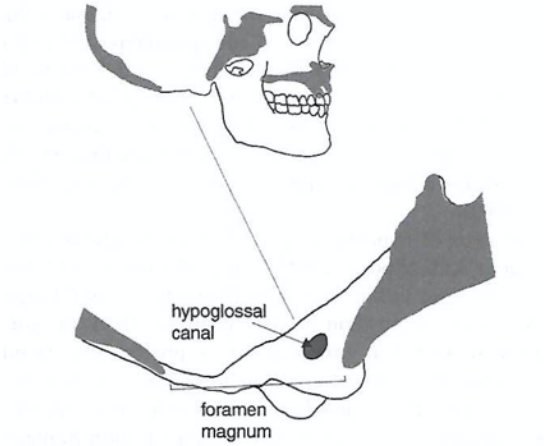

Figure 4. Location of the hypoglossal canal in the base of the human skull. Reproduced from Morley (2013: 147).

Improved control of breathing is crucial for modern human speech (de Boer 2005: 282). The size of the thoracic vertebral canal (Fig. 5) relates to the amount of thoracic innervation and is pivotal in the control of breathing (MacLarnon & Hewitt 1999). MacLarnon and Hewitt (1999: 350) argue persuasively that enhanced breathing control facilitates voice modulation, the production of long and punctuated utterances, and also "subtle variation in emphasis, pitch, and intonation". Respiratory control is critical in emphasizing particular phonemes, and thus fine control of the subglottal air pressure is important in conveying linguistic meaning (MacLarnon & Hewitt 1999: 350-351; cf. Lieberman 1984). According to MacLarnon and Hewitt (1999: 343-347), australopithecines and *Homo ergaster* have thoracic vertebral canals comparable to those of nonhuman primates, whereas the Neanderthals, early *Homo sapiens*, and contemporary humans have enlarged canals. Hence, it appears that during the long transition from *Homo ergaster* to *Homo heidelbergensis* humans acquired the ability for sophisticated respiratory control which enabled them to produce extended vocalizations and control the volume, pitch, and contour in their speech. In contrast, nonhuman primates, who lack the ability for fine breathing control, produce "only single

sound units" in short exhalations or inhalations, unable to volitionally control the duration or manipulate the pitch in their vocalizations (MacLarnon & Hewitt 1999: 353-354).

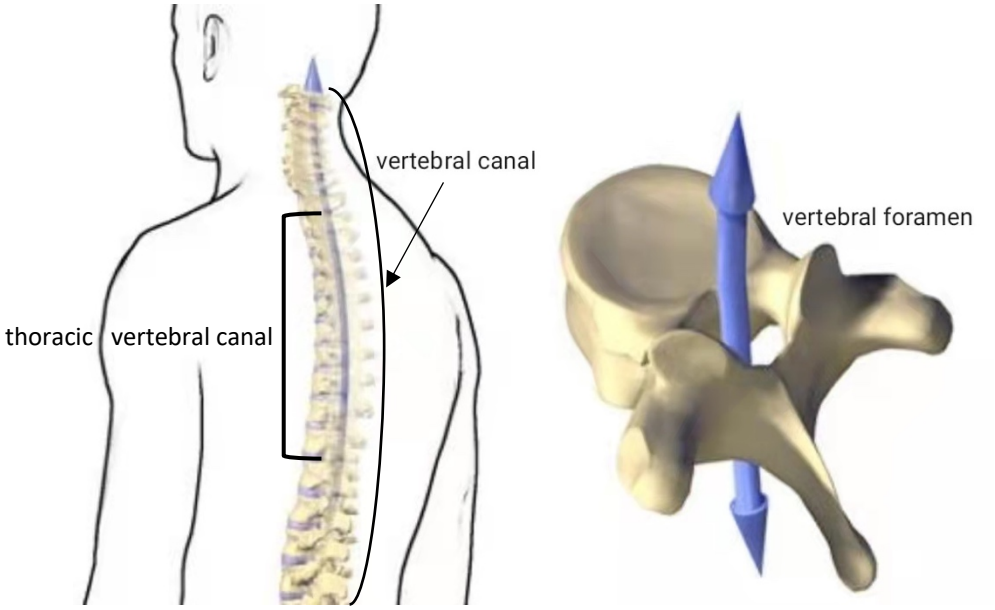

Figure 5. The vertebral canal is the hollow tubular space created by the collective stacking of the vertebral foramina, through which the spinal cord passes. An expanded canal in the area of the thorax indicates higher innervation of the muscles that control breathing. Adapted from Singh (n.d.).

The frequency range of vocalizations of a given species is tightly linked to perceptual acuity in these frequencies; this means that "the vocal and auditory systems must have evolved in response to each others' [sic] capabilities" (Morley 2013: 172). An important difference between human and chimpanzee auditory ability is that humans show a heightened sensitivity from 2,000 Hz to 4,000 Hz, whereas chimpanzees exhibit "a steep loss in sensitivity" in this frequency range (Martinez et al. 2004: 9979-9980), which is considered critical in perceiving acoustic information in human speech (Martinez et al. 2004: 9976; Dunbar 2017: 210). Martinez et al. (2004) examined the ear canal skeletal anatomy of *Homo heidelbergensis* specimens that are roughly 500,000 years old, and they conclude that their ear anatomy is essentially the same with that of modern humans, suggesting that these hominins had already developed an auditory system adapted to the frequencies of modern human speech.

The archeological record shows that several anatomical adaptations that are potentially linked to the evolution of speech had occurred by 400,000 - 600,000 years ago. Although no single piece of evidence goes unchallenged, all lines of evidence

converge to *Homo heidelbergensis*, the common ancestor of the Neanderthals and *Homo sapiens* (de Boer 2017: 159; Dunbar 2017: 210). *Homo heidelbergensis* had evolved modern-like speech production and perception anatomical adaptations, whereas australopithecines and early *Homo* did not differ significantly from extant apes. Therefore, it appears that all the necessary changes for speech occurred over a long transitional period of more than 1 million years, between *Homo ergaster* and *Homo heidelbergensis* (Dediu & Levinson 2013: 7; Morley 2013: 177). The anatomical evidence tells us when fine vocal control and the ability to produce complex vocalizations may have evolved. Although this does not entail that modern-like linguistic abilities had already evolved by 400,000 - 600,000 years ago, it shows that the modern capacity for speech (which includes sophisticated voice modulation and vocal control) had already been present at that time.

5.2.2. Paleogenetic data

Substantial developments in molecular genetics during the past two decades offer previously unimagined ways to empirically investigate the evolution of language (Fisher 2017: 34; Fitch 2017: 18). Several language-relevant genes have been identified in the human genome (Mozzi et al. 2016; Boeckx 2017: 196-197), with the gene FOXP2 having received the most attention. Comparative analyses of the DNA sequence of a language-relevant gene in humans and other species make it possible to trace its evolutionary history (Fisher 2017: 36). Importantly, the entire genomes of two extinct hominin species, namely Neanderthals and Denisovans (see Fig. 3), have been successfully extracted and sequenced, opening more ways for assessing the relevance of changes in genes such as FOXP2 (Fisher 2016: 37-38; Fitch 2017: 18).

Although FOXP2 is not the only gene involved in language, it is considered to have a pivotal role (Mithen 2005: 249-250; Dediu & Levinson 2013: 4; Fisher 2017). In humans, this gene is primary in oral motor control (Fitch 2017: 19). Nevertheless, FOXP2 is not an exclusive trait of humans; it is observed in similar form in many unrelated species (Fisher 2017: 36). In bird species that learn their vocalizations, such as in songbirds, it is expressed in regions of the brain implicated in motor control of the learned songs (Patel 2008: 389, 2014: 5). That the gene has been highly conserved

among distantly related species points to its ancient role and importance for sensorimotor functions (Fisher 2017: 36-37).

Compared to chimpanzees, the FOXP2 DNA sequence in modern humans differs in two points, which is compatible with the hypothesis that these genetic changes are relevant to language (Krause et al. 2007: 1908). Although these two mutations were originally assumed to be unique in *Homo sapiens*, Krause et al. (2007) have demonstrated that this is not the case. Their analysis of ancient DNA from Neanderthal specimens suggests that these mutations occurred 300,000 - 400,000 years ago, and possibly even earlier (Krause et al. 2007: 1910-1911).

SRGAP2 is another gene pertinent to the evolution of language. SRGAP2 has not changed during hominin evolution, but "it has given rise to three human-specific duplications, two of which underwent subsequent mutations" (Martins, Mari & Boeckx 2018: 69). *Gene duplication* is an important driver in evolution: by creating an abundance of copies of a particular gene, it provides the basis for subsequent divergence of the duplicates (see Fitch 2017: 19). The mutations of the duplicated SRGAP2 genes that are found in modern humans have also been detected in the genomes of Neanderthals and Denisovans (Martins, Mari & Boeckx 2018: 69). Crucially, the SRGAP2 duplications and their subsequent mutations are linked to the ability for vocal learning, providing further support to the growing evidence that vocal learning predates the emergence of *Homo sapiens* (Martins, Mari & Boeckx 2018: 71-75).

Turning back to the human FOXP2, it has been documented that a defective mutation in this gene causes inheritable language-related disorders, including orofacial coordination problems, difficulties in telling apart real words from nonce words, and trouble with articulation (Alcock et al. 2000a; Patel 2008: 388). Moreover, affected individuals in a study by Alcock et al. were found to have an impaired ability "in both the perception and production of rhythm in both vocal and manual modalities" (Alcock et al. 2000b: 34), whereas their pitch and melody discrimination and production performance did not differ from control subjects (Alcock et al. 2000b: 42). Alcock et al.'s (2000a, 200b) studies make clear that the impaired individuals do not have a deficit that is exclusive to linguistic processing. Rather, the problem seems to be on patterns of timing, which implicates that "fine-grained temporal processing is fundamental to speech and language" (Alcock et al. 2000b: 34). This illustrates that FOXP2 interacts with

other genes in complex neural circuits and is involved not only in linguistic processing, but also in motor control and sequencing (see Patel 2008: 389). This is also consistent with the hypothesis that complex vocal learning and the ability to synchronize to a beat are closely related (Patel 2006).

The FOXP2 and SRGAP2 mutations suggest that human species prior to *Homo sapiens* were capable of oral motor control, articulated speech, rhythm processing, and vocal learning. Thus, the studies on the language-related genes discussed here indicate that earlier human species possessed some form of language, although modern human language as we know it today may have been a more recent phenomenon.

5.3. Evolutionary steps in the emergence of language

The interdisciplinary findings discussed so far provide the basis to construct a plausible evolutionary scenario of the emergence of language and musicality. Although any attempt to do so is speculative, what can be hypothesized with some degree of confidence is the order in which certain traits appeared and the latest possible date of their appearance. It is argued here that increased vocal control and an increased range of pitch contour (see Morley 2013: 319) played a prominent role in the early stages of language evolution, followed by the capacity for vocal learning and elaborate temporal processing (cf. Alcock et al. 2000b; de Boer 2017: 160).

The first important steps in the long transition for physiological (vocal and auditory) and cognitive adaptations to speech seem to have occurred in the early *Homo* species. The appearance of *Homo ergaster* around 1.8 MYA coincides with a substantially increased brain size which has followed a continuous upward trend until the emergence of anatomically modern humans (Mithen 2005: 159; Dunbar 2016: 21). The cranial morphology and vocal anatomy of *Homo ergaster* indicate that there was a shift underway from ape-like vocalizations to the capacity for an increasing range of producible sounds (Morley 2013: 319-320). *Homo ergaster* exhibited an improving capacity for voice modulation and had a degree of vocal control not previously found in other hominins, but the current findings suggest that they were not yet able to produce extended vocalizations in the same way we do, since they lacked the capacity for sophisticated breath control (MacLarnon & Hewitt 1999). Nevertheless, *Homo ergaster*

possessed some articulatory versatility and was biologically equipped "to produce vocalizations controlled for pitch, intensity, and contour" (Morley 2013: 152). The first major anatomical and neural adaptations to speech kicked in during the rise of *Homo ergaster*, but the fossil record shows that it took more than a million years for modern-like speech capabilities to evolve. Therefore, the long period from the rise of *Homo ergaster* (ca. 1.8 MYA) to their replacement by archaic humans (starting around 600,000 years ago) may be regarded as a key transitional phase in the evolution of speech and language (cf. Dunbar 2016: 183). While it is unclear whether *Homo ergaster* had the capacity for vocal learning, the paleontological evidence shows that they possessed an increasing ability for vocal control, meaning that some of the subcomponents of language started to evolve as early as 1.8 MYA.

The fossil record reviewed in section 5.2.1 indicates that by 400,000 - 600,000 years ago archaic humans had evolved all the necessary anatomical and neurological adaptations needed for modern speech. In contrast to the preceding *Homo ergaster* populations, it seems that archaic humans had evolved the ability to produce extended vocalizations (e.g., Morley 2013: 321). Extended vocalizations are a prerequisite for language; they require prolonged exhalations, which allow for an improved control of volume, enabling us to volitionally regulate intonation and stress some parts of an utterance over others (see MacLarnon & Hewitt 1999: 358). Additionally, the paleogenetic data also suggest that by 400,000 years ago (and possibly much earlier) the capacity for vocal learning and perception of timing patterns had evolved in our lineage (Martins, Mari & Boeckx 2018; see also Alcock et al. 2000b; Krause et al. 2007). These implicate that archaic humans were biologically equipped for modern speech. While early *Homo* displayed an increasing ability for timbre and pitch modulation (cf. Fenk-Oczlon 2017: 4), archaic humans had developed the ability for more flexible vocal control, which allowed them to produce long utterances and use a broader range of frequencies.

Does this mean that archaic humans had linguistic abilities equal to our own? This is a contentious matter. Dediu and Levinson (2013: 9-12) are inclined towards a positive answer. Notwithstanding the minor biological discrepancies between *Homo heidelbergensis* and *Homo sapiens*, they place emphasis on cultural evolution processes during the last 500,000 years (Dediu & Levinson 2013: 9). Fitch (2017: 21), on the other hand, proposes that subtle genetic differences between Neanderthals and *Homo*

sapiens in regulatory elements that interact with the FOXP2 gene (see Maricic et al. 2013) may have had a critical impact on the evolution of language, pointing to the uniqueness of "full modern language" in *Homo sapiens*.

In any case, the current evidence strongly suggests that vocal control and vocal learning evolved earlier than syntactic structures and compositional meaning, which depend on intonational cues to transmit the intended information (e.g., Morley 2013: 311; see also de Boer 2017: 160-161). Regardless of whether modern, compositional language evolved in *Homo heidelbergensis* or in *Homo sapiens*, the sum of interdisciplinary findings indicates that some components of language (i.e., fine control of pitch, timbre, intensity, and duration of vocal output) evolved rather early in the human lineage. Most scholars would agree that modern linguistic skills characterized by referentiality, symbolism, propositional meaning, and hierarchical structures emerged rather late (e.g., Tomasello 2003: 11-12; Christiansen & Chater 2008: 493; Chomsky 2010: 58-59; Tattersall 2017). The same could be said for modern capacity for music, which is contingent on fine-grained rhythm perception, entrainment to an external pulse, melodic contour processing, and metrical hierarchy (see Patel 2008; Morley 2013: 321-325). I am not aware of any findings that indicate that the cognitive abilities required for hierarchical embedding evolved early in the hominin lineage. This supports the notion that a prosodic/musical system of communication preceded syntactic compositional language, which, as modern music, is marked by generativity, complex structures, and combinatorial meaning (Patel 2008; Merker, Morley & Zuidema 2015: 3; de Boer 2017: 160).

In sum, a large volume of cross-disciplinary data points to the following evolutionary steps that led to language and music as we know them today. Early *Homo* started to develop enhanced voice modulation abilities which, in the long run, resulted in flexible vocal control, forming the basis for articulated speech (Pisanski et al. 2016). Assuming that human speech evolved phylogenetically from primate vocalizations, it is reasonable to propose that in its early stages it was holistic in nature (see section 3.3), and it had a strongly affective component, relying primarily on the modulation of pitch, timbre, and amplitude (e.g., Dissanayake 2000; Morley 2013: 168, 311; Brown 2017). Thus, a prosodic protolanguage built on emotional prosody gradually evolved into a more sophisticated system of communication in which linguistic prosody was deployed

and conveyed more nuanced meanings besides emotive information (see Mithen 2005; Filippi 2016; Brown 2017). The abilities for vocal learning and audiomotor entrainment, which may depend on shared neural substrates (Patel 2014), had far-reaching effects. The ability to imitate vocal (and perhaps manual) signals gradually improved (Donald 2017), producing increasingly larger and complex vocal repertoires. This mimetic ability, along with an enhanced capacity for innovation, eventually resulted through the process of cultural evolution in the complexities of language and music described above.

6. Conclusion

The evolution of language is a complex and multi-faceted issue. A deep understanding of the question of the origins of language requires an assemblage of findings from different disciplines. Language and music have been generally assumed to be two distinct and unrelated cognitive domains, but an in-depth comparison reveals links and commonalities that may not be evident at first glance. Moreover, insights from research on prelinguistic infants, nonhuman primates, human cognitive processing, paleoanthropology, and genetics support the notion of a common precursor of language and music. Recognizing their intricate nature, and that their emergence has been the outcome of a multi-step process, is crucial in this endeavor.

The facility in which humans process relative pitch indicates that auditory processing mechanisms have been shaped by natural selection to process prosodic modulation of vocal signals (see Patel 2008: 396-398). Additionally, the archeological record suggests that the ability to control pitch, contour, and amplitude in vocalizations evolved first, followed by the ability to produce extended vocal sequences (see Morley 2013: 150-152). This points to the ancient provenance of prosody in hominin communication. The multiple genes that are implicated in language and the different variants of language-related genes indicate that earlier *Homo* species possessed oral motor control but not the full package for modern language. These suggest that the capacity for voice modulation and complex vocalizations preceded the appearance of syntax, which was a late addition in the evolution of language; even in modern language, syntactic structures occur in association with prosodic variation to convey the intended meaning. The putative protolanguage was probably holistic in nature (at least in its early

stages): the capacity to modulate the voice in a volitional and flexible manner originally conveyed emotional states and intentions holistically, before the advent of lexical items and grammar led to the analytic language we use today.

The biological foundations for language evolved gradually, possibly spanning more than a million years. Therefore, it is reasonable to proffer that the evolution of language has been subject to different selective pressures. Sociality is a long-established characteristic in the primate lineage, and thus social selective pressures could have played a key role in the emergence of language and music. Those who were skilled at modulating their vocal signals to manipulate the behavior of conspecifics would have earned a selective advantage. The growing social complexity would have contributed to a progressively expanding vocal repertoire, paving the way for modern linguistic complexity. In addition, vocal exchanges among close kin, such as mother-infant interactions, may have also been relevant.

In conclusion, explaining how language originated is a demanding task that requires interdisciplinary effort. It has been traditionally thought of as highly speculative and unscientific, but the accumulation of evidence from multiple disciplines allows the formulation of empirically driven hypotheses. The idea of a prosodic protolanguage draws on scientific evidence from a wide range of fields and poses a serious challenge to entrenched beliefs about language evolution. Addressing the selective pressures responsible for the emergence of language is more challenging, yet the relevant data point to complex social structure as a determining factor. Modern language and music in their current form might be relatively recent phenomena of learned cultural behavior, but the capacity for the perception and production of voluntary vocalizations that regulate pitch, timbre, duration, and intensity seems to have been far older.

7. References

Acosta, Jaime C.; Ward, Nigel G. 2011. “Achieving rapport with turn-by-turn, user-responsive emotional coloring”. *Speech Communication* 53, 1137-1148.

Alcock, Katherine J.; Passingham, Richard E.; Watkins, Kate; Vargha-Khadem, Faraneh. 2000a. “Oral dyspraxia in inherited speech and language impairment and acquired dysphasia”. *Brain and Language* 75, 17-33.

Alcock, Katherine J.; Passingham, Richard E.; Watkins, Kate; Vargha-Khadem, Faraneh. 2000b. “Pitch and timing abilities in inherited speech and language impairment”. *Brain and Language* 75, 34-36.

Ayala, Francisco Jose. 2019. “sexual selection”. *Encyclopedia Britannica*. https://www.britannica.com/science/sexual-selection. (23 Nov. 2021)

Beckman, Mary E. 1993. “Evidence for speech rhythm across languages”. In Tohkura, Yoh’ichi; Vatikiotis-Bateson, Eric; Sagisaka, Yoshinori (eds.). *Speech perception, production and linguistic structure*. Tokyo: IOS Press, 457-463.

Berwick, Robert C.; Okanoya, Kazuo; Beckers, Gabriel J.L.; Bolhuis, Johan J. 2011. “Songs to syntax: the linguistics of birdsong”. *Trends in Cognitive Sciences* 15(3), 113-121.

Bickerton, Derek. 2000. “How protolanguage became language”. In Knight, Chris; Studdert-Kennedy, Michael; Hurford, James (eds.). *The evolutionary emergence of language: social function and the origins of linguistic form*. Cambridge: Cambridge University Press, 264-284.

Boeckx, Cedric. 2017. “The language-ready head: evolutionary considerations”. *Psychonomic Bulletin and Review* 24, 194-199.

Botha, Rudolf. 2009. “On musilanguage/‘Hmmmmm’ as an evolutionary precursor to language”. *Language and Communication* 29, 61-76.

Brandt, Anthony; Gebrian, Molly; Slevc, Robert L. 2012. “Music and early language acquisition”. *Frontiers in Psychology* 3(327), 1-17.

Britannica, The Editors of Encyclopaedia. 2018. “kin selection”. *Encyclopedia Britannica*. https://www.britannica.com/topic/kin-selection. (23 Nov. 2021)

Brown, Steven. 2000. “The ‘musilanguage’ model of music evolution”. In Wallin, Nils L.; Merker, Bjorn; Brown, Steven (eds.). *The origins of music*. Cambridge, MA: MIT Press, 271-300.

Brown, Steven. 2017. “A joint prosodic origin of language and music”. *Frontiers in Psychology* 8(1894), 1-20.

Brown, Steven; Martinez, Michael J.; Parsons, Lawrence M. 2006a. “Music and language side by side in the brain: a PET study of the generation of melodies and sentences”. *European Journal of Neuroscience* 23, 2791-2803.

Brown, Steven; Martinez, Michael J.; Parsons, Lawrence M. 2006b. “The neural basis of human dance”. *Cerebral Cortex* 16(8), 1157-1167.

Brust, John C. M. 2003. “Music and the neurologist: a historical perspective”. In Peretz, Isabelle; Zatorre, Robert J. (eds.). *The cognitive neuroscience of music*. Oxford: Oxford University Press, 181-191.

Chomsky, Noam. 2007. “Biolinguistic explorations: design, development, evolution”. *International Journal of Philosophical Studies* 15(1), 1-21.

Chomsky, Noam. 2010. “Some simple evo devo theses: how true might they be for language?” In Larson, Richard K; Deprez, Viviane; Yamakido, Hiroko (ed/s.). *The evolution of human language*. Cambridge, UK: Cambridge University Press, 45–62.

Christiansen, Morten H.; Chater, Nick. 2008. “Language as shaped by the brain”. *Behavioral and Brain Sciences* 31, 489-558.

Corballis, Michael C. 2015. “What’s left in language? Beyond the classical model”. *Annals of the New York Academy of Sciences* 1359, 14-29.

Clay, Zanna; Archbold, Jahmaira; Zuberbuhler, Klaus. 2015. “Functional flexibility in wild bonobo vocal behaviour”. *PeerJ* 2015(8), 1-20.

Darwin, Charles. 1871. *The descent of man and selection in relation to sex*. New York: Appleton.

De Boer, Bart. 2005. “Evolution of speech and its acquisition”. *Adaptive Behavior* 13(4), 281-292.

De Boer, Bart. 2017. “Evolution of speech and evolution of language”. *Psychonomic Bulletin and Review* 24, 158-162.

Dediu, Dan; Levinson Stephen C. 2013. “On the antiquity of language: the reinterpretation of Neandertal linguistic capacities and its consequences”. *Frontiers in Psychology* 4(397), 1-17.

Dissanayake, Ellen. 2000. “Antecedents of the temporal arts in early mother-infant interaction”. In Wallin, Nils L.; Merker, Bjorn; Brown, Steven (eds.). *The origins of music*. Cambridge, MA: MIT Press, 389-410.

Donald, Merlin. 2017. “Key cognitive preconditions for the evolution of language”. *Psychonomic Bulletin and Review* 24, 204-208.

Dunbar, Robin I. M. 2016. *Human evolution*. Oxford: Oxford University Press.

Dunbar, Robin I. M. 2017. “Group size, vocal grooming and the origins of language”. *Psychonomic Bulletin and Review* 24, 209-212.

Everett, Daniel. 2017. *How language began: the story of humanity's greatest invention*. London: Profile Books.

Falk, Dean. 2000. “Hominid brain evolution and the origins of music”. In Wallin, Nils L.; Merker, Bjorn; Brown, Steven (eds.). *The origins of music*. Cambridge, MA: MIT Press, 197-216.

Falk, Dean. 2004. “Prelinguistic evolution in early hominins: whence motherese”. *Behavioral and Brain Sciences* 27, 491-541.

Farnsworth, Bryn. 2019. “EEG vs. MRI vs. fMRI – what are the differences?”. imotions.com/blog/eeg-vs-mri-vs-fmri-differences/ (17 Oct. 2021)

Fenk-Oczlon, Gertraud. 2017. “What vowels can tell us about the evolution of music”. *Frontiers in Psychology* 8(1581), 1-6.

Fenk-Oczlon, Gertraud; Fenk, August. 2009. “Some parallels between language and music from a cognitive and evolutionary perspective”. *Musicae Scientiae* 13(2), 201-226.

Fernald, Anne. 1989. “Intonation and communicative intent in mothers’ speech to infants: is the melody the message? *Child Development* 60(6), 1497-1510.

Fisher, Simon E. 2017. “Evolution of language: lessons from the genome”. *Psychonomic Bulletin and Review* 24, 34-40.

Fitch, Tecumseh W. 2000. “The evolution of speech: a comparative review”. *Trends in Cognitive Sciences* 4(7), 258-267.

Fitch, Tecumseh W. 2006. “The biology and evolution of music: a comparative perspective”. *Cognition* 100(1), 173-215.
Fitch, Tecumseh W. 2010. *The evolution of language*. Cambridge: Cambridge University Press.
Fitch, Tecumseh W. 2013. “Musical protolanguage: Darwin’s theory of language evolution revisited”. In Bolhuis, Johan J.; Everaert, Martin B. H. (eds.). *Birdsong, speech and language: exploring the evolution of mind and brain*. Cambridge, MA: MIT Press, 489-503.
Fitch, Tecumseh W. 2017. “Empirical approaches to the study of language evolution”. *Psychonomic Bulletin and Review* 24, 3-33.
Filippi, Piera. 2016. “Emotional and interactional prosody across animal communication systems: a comparative approach to the emergence of language”. *Frontiers in Psychology* 7(1393), 1-19.
Filippi, Piera; Gingras, Bruno; Fitch, Tecumseh W. 2014. “Pitch enhancement facilitates word learning across visual contexts”. *Frontiers in Psychology* 5(1468), 1-8.
Filippi, Piera; Ocklenburg, Sebastian; Bowling, Daniel L.; Heege, Larissa; Gunturkun, Onur; Newen, Albert; de Boer, Bart. 2017. “More than words (and faces): evidence for a Stroop effect of prosody in emotion word processing”. *Cognition and Emotion* 31(5), 879-891.
Foxton, Jessica M.; Dean, Jennifer L.; Gee, Rosemary; Peretz, Isabelle; Griffiths, Timothy D. 2004. “Characterization of deficits in pitch perception underlying ‘tone deafness’”. Brain 127(4), 801-810.
Friederici, Angela D. 2017. “Evolution of the neural language network”. *Psychonomic Bulletin and Review* 24, 41-47.
Geissmann, Thomas. 2000. “Gibbon songs and human music from an evolutionary perspective”. In Wallin, Nils L.; Merker, Bjorn; Brown, Steven (eds.). *The origins of music*. Cambridge, MA: MIT Press, 103-123.
Grahn, Jessica A.; Brett, Matthew. 2007. “Rhythm and beat perception in motor areas of the brain”. *Journal of Cognitive Neuroscience* 19(5), 893-906.
Gratier, Maya; Devouche, Emmanuel; Guellai, Bahia; Infanti, Rubia; Yilmaz, Ebru; Parlato-Oliveira, Erika. 2015. “Early development of turn-taking in vocal interaction between mothers and infants”. *Frontiers in Psychology* 6(1167), 1-10.
Hauser, Marc D.; Chomsky, Noam; Fitch, Tecumseh W. 2002. “The faculty of language: what is it, who has it, and how did it evolve? *Science* 298, 1569-1579.
Hopkins, William D.; Taglialatela, Jared P.; Leavens, David A. 2007. “Chimpanzees differentially produce novel vocalizations to capture the attention of a human”. *Animal Behaviour* 73, 281-286.
Houston, Derek. 2016. “Infant speech perception”. In Tharpe, Marie; Seewald, Richard (eds.). *Comprehensive handbook of pediatric audiology*. San Diego: Plural Publishing, 47-62.
Jackendoff, Ray. 2009. “Parallels and nonparallels between language and music”. *Music Perception* 26(3), 195-204.
Jespersen, Otto. 1922. *Language: its nature, development and origin*. New York: Norton.
Juslin, Patrik N.; Laukka, Petri. 2003. “Communication of emotions in vocal expression and music performance: different channels, same code? *Psychological Bulletin* 129(5), 770-814.

John Hopkins Medicine. n.d. “What is positron emission tomography (PET)?”. hopkinsmedicine.org/health/treatment-tests-and-therapies/positron-emission-tomography-pet/ (17 Oct. 2021)

Johnson, Elizabeth K.; Jusczyk, Peter W. 2001. “Word segmentation by 8-month-olds: when speech cues count more than statistics”. *Journal of Memory and Language* 44, 548-567.

Kay, Richard F.; Cartmill, Matt; Balow, Michelle. 1998. “The hypoglossal canal and the origin of human vocal behavior”. *Proceedings of the National Academy of Sciences* 95(9), 5417-5419.

Kirby, Simon. 2000. “Syntax without natural selection: how compositionality emerges from vocabulary in a population of learners”. In Knight, Chris; Studdert-Kennedy, Michael; Hurford, James (eds.). *The evolutionary emergence of language: social function and the origins of linguistic form*. Cambridge: Cambridge University Press, 303-323.

Kirby, Simon. 2001. “Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity”. *IEEE Journal of Evolutionary Computation* 5, 101-110.

Kirschner, Sebastian; Tomasello, Michael. 2010. “Joint music making promotes prosocial behavior in 4-year-old children”. *Evolution and Human Behavior* 31, 354-364.

Ko, Kwang Hyun. 2016. “Hominin interbreeding and the evolution of human variation”. *Journal of Biological Research-Thessaloniki* 23(17), 1-9.

Kotilahti, Kalle; Nissila, Ilkka; Nasi, Tiina; Lipiainen, Lauri; Noponen, Tommu; Merilainen, Pekka; Huotilainen, Minna; Fellman, Vineta. 2010. “Hemodynamic responses to speech and music in newborn infants”. *Human Brain Mapping* 31(4), 595-603.

Kotz, Sonja A.; Ravignani, Andrea; Fitch, Tecumseh W. 2018. “The evolution of rhythm processing”. *Trends in Cognitive Sciences* 22(10), 896-910.

Krause, Johannes; Lalueza-Fox, Carles; Orlando, Ludovic; Enard, Wolfgang; Green, Richard E.; Burbano, Hernan A.; Hublin, Jean-Jacques; Hanni, Catherine; Fortea, Javier; de la Rasilla, Marco; Bertranpetit, Jaume; Rosas, Antonio; Paabo, Svante. 2007. “The derived FOXP2 variant of modern humans was shared with Neandertals”. *Current Biology* 17, 1908-1912.

Laland, Kevin N.; Odling-Smee, John; Myles, Sean. 2010. “How culture shaped the human genome: bringing genetics and the human sciences together”. *Nature Reviews. Genetics* 11(2), 137-148.

Leongomez, Juan David; Binter, Jakub; Kubicova, Lydie; Stolarova, Petra; Klapilova, Katerina; Havlicek, Jan; Roberts, S. Craig. 2014. “Vocal modulation during courtship increases proceptivity even in naïve listeners”. *Evolution and Human Behavior* 35, 489-496.

Levinson, Stephen C. 2016. “Turn-taking in human communication: origins and implications for language processing”. *Trends in Cognitive Sciences* 20, 6-14.

Levi-Strauss, Claude. 1969 [1964]. *The raw and the cooked: introduction to a science of mythology*. New York: Harper and Row.

Lieberman, Philip. 1984. *The biology and evolution of language*. Cambridge, MA: Harvard University Press.

Liu, Fang; Patel, Aniruddh D.; Fourcin, Adrian; Stewart, Lauren. 2010. “Intonation processing in congenital amusia: discrimination, identification and imitation”. *Brain* 133, 1682-1693.

MacLarnon, Ann M.; Hewitt, Gwen P. 1999. “The evolution of human speech: the role of enhanced breathing control”. *American Journal of Physical Anthropology* 109, 341-363.

Maess, Burkhard; Koelsch, Stefan; Gunter, Thomas C.; Friederici, Angela D. 2001. “Musical syntax is processed in Broca’s area: an MEG study”. *Nature Neuroscience* 4(5), 540-545.

Maricic, Tomislav; Gunther, Viola; Georgiev, Oleg; Gehre, Sabine; Curlin, Marija; Schreiweis, Christiane; Naumann, Ronald; Burbano, Hernan A.; Meyer, Matthias; Lalueza-Fox, Carles; de la Rasilla, Marco; Rosas, Antonio; Gajovic, Srecko; Kelso, Janer; Enard, Wolfgang; Schaffner, Walter; Paabo, Svante. 2013. “A recent evolutionary change affects a regulatory element in the human FOXP2 gene”. *Molecular Biology and Evolution* 30(4), 844-852.

Marler, Peter. 2000. “Origins of music and speech: insights from animals”. In Wallin, Nils L.; Merker, Bjorn; Brown, Steven (eds.). *The origins of music*. Cambridge, MA: MIT Press, 31-48.

Martinez, Ignacio; Rosa, Maria; Arsuaga, Juan-Luis; Jarabo, Pilar; Quam, Rolf; Lorenzo, C.; Gracia, A.; Carretero, J.-M.; Bermudez de Castro, J.-M.; Carbonell, E. 2004. “Auditory capacities in middle Pleistocene humans from the Sierra de Atapuerca in Spain”. *Proceedings of the National Academy of Sciences* 101(27), 9976-9981.

Martins, Pedro Tiago; Mari, Maties; Boeckx, Cedric. 2018. “SRGAP2 and the gradual evolution of the modern human language faculty”. *Journal of Language Evolution* 3(1), 67-78.

Masataka, Nobuo. 2009. “The origins of language and the evolution of music: a comparative perspective”. *Physics of Life Reviews* 6, 11-22.

Merker, Bjorn. 2000. “Synchronous chorusing and human origins”. In Wallin, Nils L.; Merker, Bjorn; Brown, Steven (eds.). *The origins of music*. Cambridge, MA: MIT Press, 315-327.

Merker, Bjorn; Morley, Iain; Zuidema, Willem. 2015. “Five fundamental constraints on theories of the origins of music”. *Philosophical Transactions. Biological Sciences* 370(1664), 1-11.

Miller, Geoffrey. 2000. “Evolution of music through sexual selection”. In Wallin, Nils L.; Merker, Bjorn; Brown, Steven (eds.). *The origins of music*. Cambridge, MA: MIT Press, 329-360.

Miller, Geoffrey. 2001. *The mating mind: how sexual choice shaped the evolution of human nature*. London: Vintage.

Mithen, Steven. 2005. *The singing Neanderthals: the origins of music, language, mind and body*. London: Weidenfeld and Nicolson.

Morley, Iain. 2006. “Name that tune” (Review feature on the *The singing Neanderthals: the origins of music, language, mind and body* by Steven Mithen). *Cambridge Archaeological Journal* 16(1), 97-112.

Morley, Iain. 2013. *The prehistory of music: human evolution, archaeology, and the origins of musicality*. Oxford, Oxford University Press.

Mozzi, Alessandra; Forni, Diego; Clerici, Mario; Pozzoli, Uberto; Mascheretti, Sara; Guerini, Franca R.; Riva, Stefania; Bresolin, Nereo; Cagliani, Rachele; Sironi, Manuela. 2016. “The evolutionary history of genes involved in spoken and written language: beyond FOXP2”. *Scientific Reports* 6(22157), 1-11.

Nazzi, Thierry; Bertoncini, Josiane; Mehler, Jacques. 1998. “Language discrimination by newborns: toward an understanding of the role of rhythm”. *Journal of Experimental Psychology: Human Perception and Performance* 24(3), 756-766.
Nettl, Bruno. 1956. *Music in primitive culture*. Cambridge MA: Harvard University Press.
Nettl, Bruno. 2000. “An ethnomusicologist contemplates universals in musical sound and musical culture”. In Wallin, Nils L.; Merker, Bjorn; Brown, Steven (eds.). *The origins of music*. Cambridge, MA: MIT Press, 463-472.
Nygaard, Lynne C.; Herold, Debora S.; Namy, Laura L. 2009. “The semantics of prosody: acoustic and perceptual evidence of prosodic correlates to word meaning”. *Cognitive Science* 33, 127-146.
OED Online. 2021. “language, n.”. Oxford: Oxford University Press. https://www-oed-com.uaccess.univie.ac.at/view/Entry/105582?rskey=JvX6uF&result=1#eid (5 Oct. 2021)
OED Online. 2021. “music, n.”. Oxford: Oxford University Press. https://www-oed-com.uaccess.univie.ac.at/view/Entry/124108?rskey=1Bgr9v&result=1&isAdvanced=false#eid (1 Dec. 2021)
Oesch, Nathan. 2019. “Music and language in social interaction: synchrony, antiphony, and functional origins”. *Frontiers in Psychology* 10(1514), 1-13.
Patel, Aniruddh D. 2006. “Musical rhythm, linguistic rhythm, and human evolution”. *Music Perception* 24(1), 99-104.
Patel, Aniruddh D. 2008. Music, language, and the brain. Oxford: Oxford University Press.
Patel, Aniruddh D. 2014. “The evolutionary biology of musical rhythm: was Darwin wrong?”. *PLOS Biology* 12(3), 1-6.
Payne, Katharine. 2000. “The progressively changing songs of humpback whales: a window on the creative process in a wild animal”. In Wallin, Nils L.; Merker, Bjorn; Brown, Steven (eds.). *The origins of music*. Cambridge, MA: MIT Press, 135-150.
Peretz, Isabelle. 2003. “Brain specialization for music: new evidence from congenital amusia”. In Peretz, Isabelle; Zatorre, Robert J. (eds.). *The cognitive neuroscience of music*. Oxford: Oxford University Press, 192-203.
Peretz, Isabelle; Coltheart, Max. 2003. “Modularity of music processing”. *Nature Neuroscience* 6(7), 688-691.
Pinker, Steven. 1997. *How the mind works*. London: Allen Lane.
Pisanski, Katarzyna; Cartei, Valentina; McGettigan, Carolyn; Raine, Jordan; Reby, David. 2016. “Voice modulation: a window into the origins of human vocal control?”. *Trends in Cognitive Sciences* 20(4), 304-318.
Provine, Robert R. 2017. “Laughter as an approach to vocal evolution: the bipedal theory”. *Psychonomic Bulletin and Review* 24, 238-244.
Ramus, Franck. 2002. “Language discrimination by newborns: teasing apart phonotactic, rhythmic, and intonational cues”. *Annual Review of Language Acquisition* 2, 1-14.
Ravignani, Andrea; Thompson, Bill; Filippi, Piera. 2018. “The evolution of musicality: what can be learned from language evolution research?”. *Frontiers in Neuroscience* 12(20), 1-7.
Redhead, Gina; Dunbar, Robin I. M. 2013. “The functions of language: an experimental study”. *Evolutionary Psychology* 11(4), 845-854.

Richman, Bruce. 1993. “On the evolution of speech: singing as the middle term”. *Current Anthropology* 34(5), 721-722.
Ritt, Nikolaus. 2004. *Selfish sounds and linguistic evolution: a Darwinian approach to language change*. Cambridge: Cambridge University Press.
Rogers, Kara. 2016. "group selection". *Encyclopedia Britannica*. https://www.britannica.com/science/group-selection. (23 Nov. 2021)
Roth, Stephanie Clare. 2019. “What is genomic medicine”. *Journal of the Medical Library Association* 107(3), 442-448.
Russell, Jamie L.; Braccini, Stephanie; Buehler, Nicole; Kachin, Michael J.; Schapiro, Steven J.; Hopkins, William D. 2005. “Chimpanzee (pan troglodytes) intentional communication is not contingent upon food”. *Animal Cognition* 8(4), 263-272.
Saffran, Jenny R.; Aslin, Richard N.; Newport, Elissa L. 1996. “Statistical learning by 8-month-old infants”. *Science* 227(5294), 1926-1928.
Savage, Patrick E.; Brown, Steven; Sakai, Emi; Currie, Thomas E. 2015. “Statistical universals reveal the structures and functions of human music”. *Proceedings of the National Academy of Sciences* 112(29), 8987-8992.
Scott-Phillips, Thomas C. 2017. “Pragmatics and the aims of language evolution”. *Psychonomic Bulletin and Review* 24, 186-189.
Searle, John R. 2016. “Human social reality and language”. *Phenomenology and Mind* 2, 26-39.
Selkirk, Elisabeth. 1984. *Phonology and syntax: the relation between sound and structure*. Cambridge, MA: MIT Press.
Sergent, Justine; Zuck, Eric; Terriah, Sean; MacDonald, Brennan. 1992. “Distributed neural network underlying musical sight-reading and keyboard performance”. *Science* 257, 106-109.
Seyfarth, Robert M.; Cheney, Dorothy L. 2017. “Precursors to language: social cognition and pragmatic inference in primates”. *Psychonomic Bulletin and Review* 24, 79-84.
Singh, Arun Pal. n.d. “Vertebral canal anatomy and contents”. http://boneandspine.com/vertebral-canal-anatomy/ (14 Oct. 2021)
Singh, Sanjay P. 2014. “Magnetoencephalography: basic principles”. *Annals of Indian Academy of Neurology* 17(1), 107-112.
Slater, Peter. 2000. “Birdsong repertoires: their origins and use”. In Wallin, Nils L.; Merker, Bjorn; Brown, Steven (eds.). *The origins of music*. Cambridge, MA: MIT Press, 49-63.
Soderstrom, Melanie; Seidl, Amanda; Nelson, Deborah G. Kemler; Jusczyk, Peter W. 2003. “The prosodic bootstrapping of phrases: evidence from prelinguistic infants”. *Journal of Memory and Language* 49, 249-267.
Stivers, Tanya; Enfield, N. J.; Brown, Penelope; Englert, Christina; Hayashi, Makoto; Heinemann, Trine; Hoymann, Gertie; Rossano, Federico; de Ruiter, Jan Peter; Yoon, Kyung-Eun; Levinson, Stephen C. 2009. “Universals and cultural variation in turn-taking in conversation”. *Proceedings of the National Academy of Sciences* 106(26), 10587-10592.
Tallerman, Maggie. 2006. “Abracadabra! Early hominin for ‘I think my humming’s out of tune with the rest of the word’” (Review feature on the *The singing Neanderthals: the origins of music, language, mind and body* by Steven Mithen). *Cambridge Archaeological Journal* 16(1), 97-112.

Tallerman, Maggie. 2013. “Join the dots: a musical interlude in the evolution of language?”. *Journal of Linguistics* 49, 455-487.

Tattersall, Ian. 2017. “How can we detect when language emerged?”. *Psychonomic Bulletin and Review* 24, 79-84.

Tomasello, Michael. 2003*. Constructing a language: a usage-based theory of language acquisition.* Cambridge, MA: Harvard University Press.

Trehub, Sandra E. 2003. “The developmental origins of musicality”. *Nature Neuroscience* 6(7), 669-673.

Wray, Alison. 1998. “Protolanguage as holistic system for social interaction”. *Language and Communication* 18, 47-67.

Wray, Alison. 2000. “Holistic utterances in protolanguage: the link from primates to humans”. In Knight, Chris; Studdert-Kennedy, Michael; Hurford, James (eds.). *The evolutionary emergence of language: social function and the origins of linguistic form*. Cambridge: Cambridge University Press, 285-302.

Zatorre, Robert J. 2003. “Neural specializations for tonal processing”. In Peretz, Isabelle; Zatorre, Robert J. (eds.). *The cognitive neuroscience of music*. Oxford: Oxford University Press, 231-247.