Primates (2022) 63:417–428
https://doi.org/10.1007/s10329-021-00905-x

SPECIAL FEATURE: REVIEW ARTICLE

Evolutionary origins of human language

# The evolution of hierarchical structure building capacity for language and music: a bottom-up perspective

Rie Asano[1]

Received: 8 November 2020 / Accepted: 26 March 2021 / Published online: 11 April 2021


## Abstract

A central property of human language is its hierarchical structure. Humans can flexibly combine elements to build a hierarchical structure expressing rich semantics. A hierarchical structure is also considered as playing a key role in many other human cognitive domains. In music, auditory-motor events are combined into hierarchical pitch and/or rhythm structure expressing affect. How did such a hierarchical structure building capacity evolve? This paper investigates this question from a bottom-up perspective based on a set of action-related components as a shared basis underlying cognitive capacities of nonhuman primates and humans. Especially, I argue that the evolution of hierarchical structure building capacity for language and music is tractable for comparative evolutionary study once we focus on the gradual elaboration of shared brain architecture: the cortico-basal ganglia-thalamocortical circuits for hierarchical control of goal-directed action and the dorsal pathways for hierarchical internal models. I suggest that this gradual elaboration of the action-related brain architecture in the context of vocal control and tool-making went hand in hand with amplification of working memory, and made the brain ready for hierarchical structure building in language and music.

**Keywords** Comparative research · Evolution · Language · Music · Hierarchical structure building · Working memory

## Introduction

Language and music as cognitive systems form a mosaic consisting of multiple components as parts with different evolutionary origins (Fitch 2006; Boeckx 2013). From a comparative language-music perspective, some of these components might be shared and based on the same evolutionary genesis, while others might be different and emerged independently in the course of evolution. To date, several candidates for shared and distinct components have been proposed in both theoretical and empirical research (e.g., Patel 2008, 2012, 2013; Jackendoff 2009; Koelsch 2012; Peretz 2013; Asano and Boeckx 2015). One candidate shared component with the same evolutionary genesis is the capacity of hierarchical structure building. While some researchers deny the possibility of investigating the evolution of biological capacity underlying human cognitive systems by studying nonhuman animals (Hauser et al. 2014), the others emphasize a set of cognitive capacities that shed light on the evolution of uniquely human and domain-specific capacities (Boeckx 2017; Fitch 2017). This paper adopts the latter bottom-up perspective and surveys how the continuum between nonhuman primates' cognitive capacities and human hierarchical structure building capacity looks like (Wakita 2020 for a similar approach). This work builds on my earlier proposal to investigate syntax in language and music in terms of the basic action-related components such as goal, planning, control, and sensory-motor integration (Asano and Boeckx 2015) and extends to a working memory framework for comparative evolutionary research. The key idea of the current paper as well as the central brain structures are summarized in Fig. 1.

✉ Rie Asano
rie.asano@uni-koeln.de

[1] Systematic Musicology, Institute of Musicology, University of Cologne, Cologne, Germany

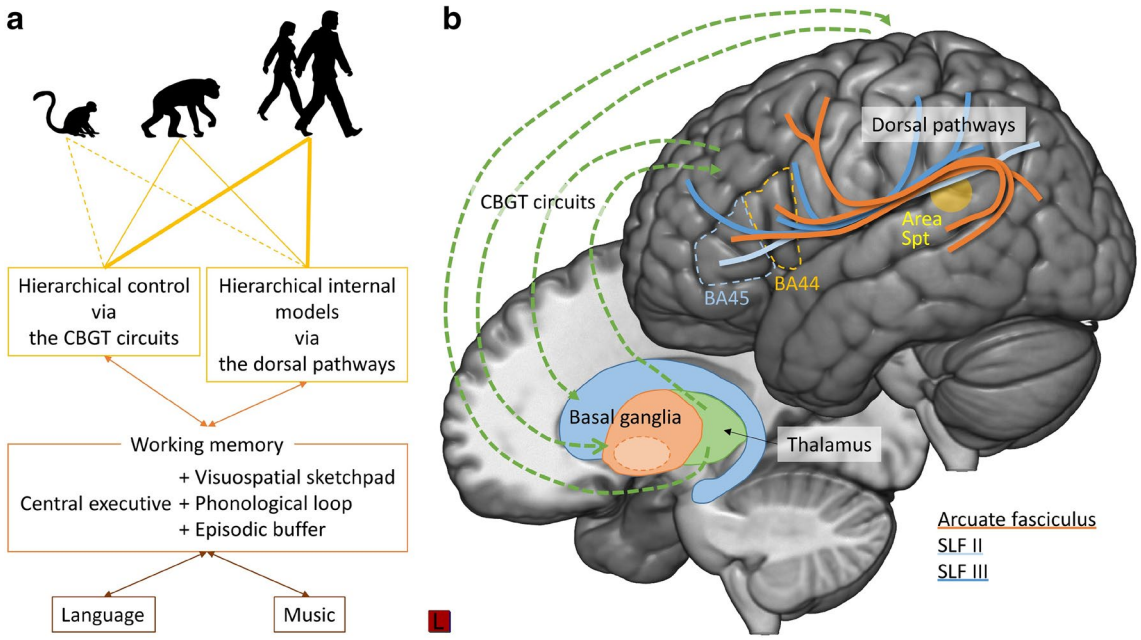


**Fig. 1** **a** Key idea of the current paper. Hierarchical control of goal-directed actions and hierarchical internal models are two central components of hierarchical structure building capacity. The basic brain architecture is shared between nonhuman primates and humans, but has gradually elaborated in the course of the evolution (yellow lines). This elaboration went hand in hand with the elaboration of working memory. Domain-specific representations of language and music emerged as a working memory resource-freeing strategy. **b** Brain structures central to the current paper. This figure displays the CBGT circuits (green dashed arrows), the dorsal pathways (arcuate fasciculus: orange line; SLF II: light blue line; SLF III: blue line), Broca's region (BA44 and BA45), and the area Spt (yellow). The brain image was created using a Montreal Neurological Institute (MNI) template provided by MRIcrogl (https://www.mccauslandcenter.sc.edu/mricrogl/). The dorsal pathways were drawn based on Petrides (2014). Abbreviations: area Sylvian parietal-temporal (area Spt); cortico-basal ganglia-thalamocortical circuits (CBGT circuits); superior longitudinal fasciculus (SLF)

## Hierarchical structure building from an action-oriented perspective

Language and music rely on three subcomponents of hierarchical structure building. The first subcomponent deals with a layered hierarchical structure with different levels of abstraction. Concerning language, phonemes combine into syllables, syllables combine into words, words combine into sentences, and sentences combine into discourses. As for music, pitch or sound events are hierarchically combined into larger units to encode affect, i.e., tension-relaxation patterns (Lerdahl and Jackendoff 1983). The second subcomponent deals with asymmetrical headed hierarchy. The linguistic syntactic structure consists of phrases with a head component determining the category of each phrase (e.g., a verb is the head of a verb phrase) and a non-head element called "complement." In musical structure, the head is the structurally most important event of a musical unit determined by rhythmic and tonal-harmonic stability, and the non-head element called "elaboration" is a less important event (Lerdahl and Jackendoff 1983). There is a considerable difference between language and music in the way how the head is determined (Jackendoff 2009; Asano and Boeckx 2015), but the fact that they are both organized asymmetrically is a crucial similarity. The third subcomponent deals with hierarchically structured long-term memory representations called "rules," "constructions," "templates," or "schemas." The long-term memory representations are domain-specific and thus can be understood as the primary source of differences between language and music (Patel 2008). They are associated with meaning in language and with affect in music.

Hierarchical structures of language and music parallel action syntax which yields flexible action organization by building asymmetrical headed hierarchical structures and layered hierarchical structures. The former aspect of action syntax deals with the hierarchical combination of preparation and goal (i.e., head) for action planning (Jackendoff 2009). The latter deals with the hierarchical combination of lower-level goals (e.g., filling a pot with water, placing a filter in a machine) into higher-level goals (e.g., filling a machine with water, filling the filter with grained coffee) to achieve the main goal (e.g., making coffee). The higher the hierarchical level is, the more abstract and temporally extended the goals become. The concept of action syntax and its relation to syntax in language and music can be traced back to the seminal work of Lashley (1951) arguing

against associative chain theories and proposing a hierarchical model of action sequencing as an alternative. He also suggested that hierarchical organization is characteristic of all skilled acts, including language and music.

Despite this striking parallel between the hierarchical structure in language, music, and action, research on the evolution of language and music, which takes the action-related components into account, rather focused on speech and beat-based timing (with an important exception of Fitch and Martins 2014). In particular, the action simulation for auditory perception (ASAP) hypothesis (Patel and Iversen 2014) and the gradual audiomotor evolution (GAE) hypothesis (Merchant and Honing 2014) have advanced this research area from an evolutionary neuroscience perspective. The ASAP hypothesis emphasizes the dorsal auditory pathway (SLF II) via the parietal cortex for beat processing and suggests that this pathway (especially its temporoparietal division of the SLF II) was enhanced in humans due to vocal learning (see also Cannon and Patel 2021 for a recent elaboration). In contrast, the GAE hypothesis claims that beat-based timing gradually evolved in primate lineage independently of vocal learning and emphasizes the role of the motor cortico-basal ganglia-thalamocortical (CBGT) circuit as a neural basis of timing shared between nonhuman primates and humans. It hypothesizes that in the course of the evolution in the primate lineage, the auditory system has gained increasingly privileged access to the motor CBGT circuit via the direct projection from the primary auditory cortex to the basal ganglia or via the dorsal auditory pathways.

The central claim which I would like to put forward here is that hierarchical structure building capacity for language and music builds on the same action-related neural architecture suggested by the ASAP hypothesis and the GAE hypothesis. The CBGT circuits and the dorsal pathways provide a foundation of hierarchical structure building: hierarchical control of goal-directed action (Badre and Nee 2018) and hierarchical internal models (Wolpert et al. 2003), respectively. In either case, hierarchy deals with different degrees of abstraction. The hierarchical organization of action implies that understanding and executing action sequences require hierarchical control, i.e., maintenance, selection, and inhibition of goals at differently abstract levels as well as the top-down control from more abstract onto more concrete levels (Badre 2008). The low-level goals directly correspond with concrete acts, while the higher-level goals are temporally extended and only indirectly govern concrete acts. The hierarchical organization of action also implies that, at multiple levels of abstraction, forward models should predict sensory consequences of actions, and inverse models should determine motor command based on desired outcomes (Wolpert et al. 2003). The lower-level internal models deal with concrete sensory consequences and motor commands, while the higher-level internal models deal with more abstract features. The interaction between hierarchical control and hierarchical internal models enables hierarchical structure building in sequencing.

Hierarchical control and hierarchical internal models are important constituents of hierarchical structure building capacity for language and music. For example, Hickok (2012) proposed the hierarchical state feedback control (HSFC) model, which investigates speech motor control as a hierarchical organization of internal models. The HSFC model includes phoneme-level, syllable-level, word-level, and conceptual-level internal models, which interact with each other. Bornkessel-Schlesewsky and colleagues (2015) suggested that a hierarchical organization of internal models can be also applied for processing linguistic sequences with various hierarchical levels including sentence-level and discourse-level. Prediction signals of the higher-level forward models propagate to the lower levels from top-down, while the lower-level error signals propagate to the higher levels to train the inverse models from bottom-up. Concerning music, Koelsch and colleagues (2019) proposed a similar hierarchical generative model of prediction based on predictive coding/active inference. Proksch and colleagues (2020) also interpreted beat-based timing from a perspective of predictive coding/active inference. As for hierarchical control, both language and music engage frontal goal hierarchies and the CBGT circuits (Ullman 2006; Kotz et al. 2009; Jeon 2014; Jeon and Friederici 2015; Lieberman 2016; Asano 2019). Thus, hierarchical control and hierarchical internal models provide a basis for hierarchical structure building in language and music.

## Building blocks for hierarchical structure building capacity in nonhuman primates

The bottom-up approach based on the action-related components makes the evolution of hierarchical structure building capacity for language and music tractable in terms of the continuity between nonhuman primates and humans. As I will argue below, nonhuman primates' auditory-motor and visuomotor systems are equipped with the basic architecture for the interplay between hierarchical control and internal models, but are not fully expanded to account for more complex hierarchical structure building seen in language and music. The limitation is more evident in their auditory-motor systems than in visuomotor systems. In the latter case, some degree of hierarchical processing exists. Moreover, I argue that research on hominin tool-making contributes to the research on how the action-related neural architecture was gradually elaborated in the hominin evolution.

## Orofacial and vocal control

Nonhuman primates possess a set of prerequisites for a component necessary for hierarchical structure building in speech (or phonology), namely the capacity for synchronized and voluntary orofacial and vocal control generating a sequential vocal stream. Flexible, modifiable, and variable orofacial and vocal behavior is present in some nonhuman primates although they do not belong to the "classical" vocal learners (for discussions on vocal learning in primates, see Lameira 2017; Fischer and Hammerschmidt 2020; Martins and Boeckx 2020).

For example, following MacNeilage's (1998) frame/content theory, Ghazanfar and Takahashi (2014a, b) point to monkey lip-smacking (facial expression including oscillatory jaw movement) and call (vocal expression) as precursors for speech and propose the synchronization between facial and vocal expression as the babbling-like first step for the emergence of speech featuring syllable-like sequences in the hominin line. They point out that the evolution of the synchronization between facial and vocal expressions as audiovisual communication signals may be simple because the gelada baboon, a unique nonhuman primate species showing such a synchronization known as a wobble, is closely related to the yellow baboon not displaying anything resembling a wobble. Recently, Risueno-Segovia and Hage (2020) showed that marmoset monkeys display synchronized jaw movement and vocalization in producing phee calls, providing evidence for early emergence of synchronized orofacial and vocal control yielding a sequential vocal stream. However, its neural mechanism remains unrevealed. In humans, Brown and colleagues (2020) determined that somatotopic representations of the larynx and jaw muscles overlap in the primary motor cortex, which could be one of the candidates making such a synchronization possible.

In addition to synchronized orofacial and vocal control, researchers identified voluntary vocal control as an ability required for generating speech-like sequential vocal streams. The ventrolateral prefrontal cortex of nonhuman primates also plays a crucial role in voluntary vocal control. In their review, Hage and Nieder (2016) provided evidence that the monkey's ventrolateral prefrontal cortex is associated with the preparation for vocalization. In particular, macaque area 44, which is comparable to human BA44, was found to be involved with the orofacial musculature (Petrides et al. 2005) and is highly connected with the larynx motor cortex (Kumar et al. 2016). Loh and colleagues (2017) reviewed the distinctive role of the ventrolateral frontal cortex, the anterior/mid-cingulate cortex, and the (pre-)supplementary motor area of humans and nonhuman primates in the voluntary vocal control: the ventrolateral frontal cortex is associated with the selection of orofacial-vocal responses based on sensory-motor conditional rules (i.e., IF Stimulus A, THEN Vocalization A); the mid-cingulate cortex is involved in evaluating consequences of the responses; and the pre-supplementary motor area is associated with the general initiation and adjustment of the responses. Thus, although preparation for vocalization is more effortful than for action execution in macaques (Koda et al. 2018), the basic mechanism underlying voluntary vocal control seems to be present.

Although some homologous mechanism is present in nonhuman primates, the human voluntary vocal control mechanism contains at least three major novelties in comparison to that of nonhuman primates. The first is the direct cortico-ambigual connection which is absent in monkeys, sparse, if at all, in chimpanzees, and enhanced in humans (Kuypers 1958a, b; Jürgens 2002). It was hypothesized to have emerged through an exaptation of the corticospinal tract subserving manual motor control (Fitch 2011). The second is the two-part structure of the human larynx motor cortex consisting of the ventral larynx cortex, which is homologous to that of monkeys and chimpanzees, and the uniquely human dorsal larynx cortex which was hypothesized to have emerged through the duplication of either vocalization-related motor areas or adjacent non-vocal motor areas (Belyk and Brown 2017). It is also the dorsal larynx motor cortex that overlaps with the jaw motor cortex (Brown et al. 2020). The third is increased connectivity of the human dorsal larynx cortex with structures in the parietal lobe including the somatosensory cortex and the inferior parietal lobe, while the structural networks of the larynx motor cortex in monkeys and humans are mainly similar otherwise (Kumar et al. 2016). Those changes may have elaborated basic mechanisms shared among humans and nonhuman primates to realize hierarchical planning and control of vocalization.

Recently, by referring to the study conducted by Kumar and colleagues (2016), Hickok (2017) argued that the dorsal stream including the area Spt (Sylvian parietal-temporal) elaborated gradually as a sensory feedback control circuit in the context of the evolution of voluntary laryngeal control. The area Spt is a crucial part of the dorsal auditory pathway and was suggested to be an auditory-motor interface system (Hickok and Poeppel 2004, 2007). Aboitiz (2017, 2018) sees the evolution of the dorsal pathways linking to the posterior Broca's region as a continuum of this elaboration. Dorsal auditory pathways—although sparse—are already present in macaques (Petrides 2014; Balezeau et al. 2020) and show a gradual expansion via chimpanzees to humans (Rilling et al. 2008; Balezeau et al. 2020). In humans, the dorsal pathways were suggested to implement the sensory-motor prediction and feedback control at differently abstract hierarchical levels (Bornkessel-Schlesewsky et al. 2015; Hickok 2017). While the auditory dorsal pathways of nonhuman primates and humans implement internal models, hierarchical levels of internal models are limited in nonhuman primates

(Bornkessel-Schlesewsky et al. 2015). Such a hierarchical organization of internal models may go hand in hand with hierarchical chunking implemented in the CBGT circuits (Rauschecker 2012). Macaques show activations in the frontal lobe, the parietal and temporal lobe, and the basal ganglia in processing auditory chunks, indicating auditory-motor integration (Uhrig et al. 2014). However, the hierarchical levels of chunks seem to be more limited in nonhuman primates than in humans (Merchant and Honing 2014). In this way, the gradual elaboration of the dorsal auditory pathway together with the CBGT circuits could have led to the hierarchical vocal control.

## Action, tool use, and tool-making

Research on planning and control provides an optimal opportunity to study the continuity of hierarchical structure building capacity in humans and nonhuman primates, as the rich capacity for action planning and control is well known in nonhuman primates. For example, Weiss and colleagues (2012) argued that cotton-top tamarin monkeys and lemurs are capable of short-span anticipatory planning as they adjust their hand posture for grabbing a cup in advance according to the affordance of the cup. Wynn and colleagues (2011) pointed out that, in addition to great apes, the long-tailed macaque and the bearded capuchin use tools to extract food that is otherwise not accessible. Although they do not discuss tool use in terms of planning, tool use as such can be regarded as involving at least one preparatory action: grasping a stone (i.e., preparation) to crack a nut (i.e., head). Byrne and colleagues (2013) reviewed evidence for chimpanzee's planning capacity concerning tool use in the wild: chimpanzees manufacture and repair tools or pick up suitable materials in advance. This indicates that chimpanzees can prepare for the main action in a more extended time range.

Action planning and control of nonhuman primates are hierarchically organized. Byrne and colleagues (2013) argued that chimpanzee's tool use is hierarchically organized because they can flexibly omit redundant steps in the tool-use sequence to achieve the goal of fishing termites. Byrne and Russon (1998) showed the hierarchical organization of a mountain gorilla's food preparation and an orangutan's imitation. They suggested that great apes can represent and manipulate the relationship between objects at different hierarchical levels, although the hierarchical depth of planning is limited. Matsuzawa (1996) showed the hierarchical organization of chimpanzees' tool-use behavior. He suggested nut-cracking behavior which includes the use of a wedge stone as the most complex form of tool use in wild chimpanzees with nested hierarchical levels. The wedge stone supports an anvil stone on which the nut is placed and hit by a hammerstone. Matsuzawa (1991) and Hayashi (2007) demonstrated that captive chimpanzees can hierarchically combine multiple cups. All those examples can be interpreted such that great apes are able to organize action into hierarchical structures consisting of preparation and the head.

Archaeological evidence suggests a gradual elaboration of action syntax in the hominin evolution. For example, Moore (2010) showed that early stone tool-making solely required serial flaking to remove high mass from the core, while hierarchical flaking with (multiple) preparatory steps is necessary for Acheulean and Lavallois tool-making. Although Stout (2011) also ascribes hierarchical organization to early stone tool-making in the Oldowan industry, it can be claimed that hierarchical organization is not necessary for Oldowan tool-making given the absence of preparatory flaking. His research shows that hierarchical flaking with a preparatory step emerged in early Acheulean tool-making and was elaborated through additional hierarchical levels in late Acheulean and Lower Palaeolithic tool-making. Stout and colleagues (2018) also showed that Acheulean tool-making is computationally more demanding than Oldowan tool-making. Ambrose (2001) even suggested that neither Oldowan nor Acheulean tool-making, but Middle Paleolithic/Middle Stone Age tool-making by Neanderthals, late archaic humans, and anatomically modern humans yielding composite tools is hierarchical. He suggested that combining technological units such as a shaft, a stone insert, and binding materials in different configurations generates functionally different tools.

Evidence from paleoneurology investigating hominin endocranial morphology suggests a series of neuroanatomical changes. Holloway (2015) suggests the relative expansion of the posterior parietal association cortex at the cost of the visual cortex approximately from 3.5 to 3.0 million years ago (Australopithecus) and the reorganization of the Broca's region approximately 1.8 million years ago (*Homo rudolfensis*). Bruner (2004, 2010) further suggests the widening of the frontal cortex as one change that happened in Neanderthals and Homo sapiens, but globularity and the general enlargement of the entire parietal surface as traits unique to Homo sapiens and absent in adult Neanderthals. Globularity is associated with parietal and cerebellar bulging (Neubauer et al. 2018).

Although they are general neuroanatomical changes and cannot be directly linked to the gradual elaboration of action syntax, there is at least evidence that some of those structures change in the context of tool use and tool-making. The parietal cortex shows tool-use-induced expansion in monkeys (Quallo et al. 2009). There is a heritable link between the individual variation in tool-use capacity and the morphology in temporal, parietal, and cerebellar cortices of chimpanzees (Hopkins et al. 2019). Cerebellum size correlates more strongly with foraging skills than social group size, while neocortex size shows

the reverse correlation pattern (Barton 2012). Moreover, Hecht and colleagues (2015b) showed that Paleolithic tool-making training remodels frontoparietal circuits in humans. Neuroimaging studies investigating stone tool-making in humans showed that the posterior parietal region plays a significant role in sensory-motor integration required for Oldowan flaking (Stout and Chaminade 2007; Stout et al. 2008). Additional right inferior frontal gyrus activation is significant for hierarchical control required for Acheulean flaking (Stout et al. 2008, 2011; Putt et al. 2017).

In humans, hierarchical control recruits the frontal lobe, the parietal lobe, the basal ganglia, and the cerebellum (Balleine et al. 2015; Badre and Nee 2018; Badre and Desrochers 2019; D'Mello et al. 2020). Both in humans and nonhuman primates, the frontal lobe is organized to form the rostrocaudal abstraction gradient: the more caudal region processes more concrete sensory-motor representation, while the more rostral region more abstract representation (Badre and D'Esposito 2009). The frontal lobe of humans and nonhuman primates projects massively to the basal ganglia which play a crucial role in goal-directed (i.e., reward-based) control and learning (Alexander et al. 1986; Middleton and Strick 2000b; Haber 2003, 2016). The cerebellum of humans and nonhuman primates works in tandem with the frontal lobe, the basal ganglia, and the parietal lobe, and plays a crucial role in building internal model and optimizing behavior (Wolpert et al. 1998; Middleton and Strick 2000a; Ramnani 2006; Ito 2008; Bostan and Strick 2018). Thus, the basic brain architecture underlying action syntax is the same for humans and nonhuman primates.

One prominent difference in the hierarchical control network is one of the frontoparietal circuits, namely the superior longitudinal fasciculus (SLF III), which is right-lateralized and larger, and projects more strongly to the inferior frontal gyrus in humans than in chimpanzees (Hecht et al. 2015a). Fronto-parieto-temporal connections via the superior and middle longitudinal fasciculi are more dominant in humans, while frontotemporal connections via the extreme/external capsules are more prominent in macaques, and the fronto-parieto-temporal connectivity profile in chimpanzees is intermediate (Hecht et al. 2013). The SLF III as a dorsal pathway enables integration of complex goals represented in the frontal lobe and increasingly complex sensory-motor representations in the parietal lobe (Stout and Hecht 2017), which supports increasingly complex hierarchical structure building through hierarchical internal models (see Wolpert et al. 2003 for hierarchical internal models). In this way, the gradual elaboration of the SLF III could have led to the amplification of action syntax.

## Integration into a working memory framework for comparative evolutionary study

In Sects. 2 and 3, I argued that the evolution of hierarchical structure building capacity for language and music can be studied based on the action-related components, and the basic brain architecture for the action-related components is shared between humans and nonhuman primates. In the current section, I suggest that the gradual elaboration of the action-related brain architecture in the context of vocal control and tool-making went hand in hand with amplification of working memory, and made the brain ready for hierarchical structure building in language and music.

Working memory is not a passive store, but an active information processing device consisting of multiple components including central executive, visuospatial sketchpad, phonological loop, and episodic buffer (for reviews, see Baddeley 2010, 2012). The central executive is an attentional control system that is supported by two short-term maintenance systems: the visuospatial sketchpad for visual information and the phonological loop for verbal and acoustic information. Each of them contains two components: a store (also buffer) that records (sensory) memory traces and a (covert motor) rehearsal process that refreshes and maintains those traces. That is, both visuospatial sketchpad and phonological loop can be conceived as interfaces for sensory-motor integration. The episodic buffer is a short-term store holding multimodal memory units, provides a platform where different components of working memory interact, and interfaces with long-term memory. Importantly, working memory is a fluid system requiring only temporary activation, while long-term memory is a crystallized system representing skills and knowledge.

Gradually increasing complexity in tool-making went hand in hand with the elaboration of the visuospatial sketchpad and the central executive. For example, Oldowan flaking requires sophisticated visuomotor coordination through more than a few hours of training (Stout 2010; Stout and Chaminade 2012), and its accuracy differs between experts and novices (Bril et al. 2010; Nonaka et al. 2010). That is, the integration of visuospatial and motor representations through the visuospatial sketchpad is central to Oldowan tool-making. The maintenance of stable visual representations through the visuospatial sketchpad is crucial for tool-making in general in learning to associate motor control parameters and visual consequences (Coolidge and Wynn 2005). Moreover, the emergence of explicit preparatory steps, which can be embedded in another preparatory action, in Acheulean tool-making by *Homo erectus* (Moore 2010; Stout 2011)

implies a need for more sophisticated planning ability as well as the ability to maintain and manipulate temporarily extended goal and subgoal representations in the mind. Finally, Middle Paleolithic/Middle Stone Age composite tool-making requires long-range planning and coordination of multiple task sets (Ambrose 2001). That is, the central executive gradually elaborated to control, i.e., select and inhibit, representations maintained at multiple hierarchical levels and in multiple task sets.

Further, I propose that the homologous neural circuits for auditory-vocal-orofacial control present in nonhuman primates provide a basis for the phonological loop. Although there is evidence that auditory long-term and short-term memory are limited in monkeys (e.g., Fritz et al. 2005; Scott et al. 2012), a recent review points to the similarity of auditory memory storage in humans and nonhuman primates (Scott and Mishkin 2016). Moreover, nonhuman primates can process auditory sequences with nonadjacent dependency (e.g., $AB^nA$) which requires a long-term memory representation of a canonical auditory pattern and an auditory storage capacity to hold an element over one or more intervening elements and compare it to another element (for reviews, Wilson et al. 2017, 2020; Petkov and ten Cate 2020). Thus, the emphasis was put on the rehearsal mechanism mapping sounds to articulatory movement via the elaborated dorsal auditory pathways rather than on the auditory memory (Aboitiz et al. 2006, 2010; Scott and Mishkin 2016; Aboitiz 2017, 2018). The phonological loop as an auditory-motor loop is based on an internal model: a forward model from the frontal cortex to the temporal cortex predicts the auditory consequence of an action, an inverse model from the temporal cortex to the frontal cortex transforms desired auditory outcome into motor command, and the area Spt serves as an auditory-motor interface (Buchsbaum and D'Esposito 2019).

Aboitiz and his colleagues argued that an enhanced capacity of the phonological loop in humans was a key to learn and process complex hierarchical sequences with multiple embeddings (Aboitiz et al. 2006, 2010). Coolidge and Wynn (2007) made a similar suggestion. However, the increased maintenance capacity alone does not lead to hierarchical structure nor embedding. Alternatively, I suggest that hierarchical structure relates rather to a memory-optimization strategy. As a short-term memory component, the phonological loop has a limited maintenance capacity of approximately three to five elements (Cowan 2010). Thus, this limited capacity should be optimally used in processing sequences. Chunking, i.e., grouping multiple elements as a unit, is a strategy to optimize memory resource use and can be regarded as compression of information at multiple hierarchical levels (Snyder 2016; Christiansen and Chater 2016). The higher the levels of chunks are, the more abstract the representations become. Rules, then, represent the (hierarchical) relationship between chunks and allow for more complexity in sequencing once stored in the long-term memory. Petkov and Wilson (2012) proposed that selective pressures to reduce memory demands through rule-based learning strategies may have expanded human hierarchical structure building capacity. Coolidge and Wynn (2005) point out the importance of learned rules stored in long-term memory for freeing up working memory capacity to access and manipulate specific content.

In comparative language-music research, working memory was suggested as a candidate shared mechanism underlying hierarchical structure building in language and music, as it is required for maintaining and manipulating intermediate results (Kljajevic 2010; Fitch and Martins 2014). Building up multiple parallel hierarchical structures, hierarchical structures with multiple embedding, or hierarchical structures deviating from canonical long-term memory representations leads to interference effects between language and music processing as extraordinary demand is placed on the working memory resources. Once units and rules are stored in the long-term memory and hierarchical processing can be partially automatized, the processing load decreases. Thus, freeing of working memory resources is one possible reason for functional specialization within the shared brain architecture for different cognitive systems.

How does functional specialization emerge? This question, then, can be tackled within a domain-*relevant* approach. Its main idea is that through neural competition, brain networks become relatively domain-specific overtime (Karmiloff-Smith 2013). That is, specialization of function can be regarded as fine-tuning of coarsely coded systems with domain-relevant biases. For example, the CBGT circuits implement hierarchical structure building in language and music through specialized parallel subcircuits (Ullman 2006; Asano 2019). The dorsal auditory pathways also show specialization such that the left arcuate fasciculus plays a more important role for hierarchical structure building in language than in music, while the reverse is true for the right arcuate fasciculus (Friederici 2019). In this way, language and music can be regarded as different cognitive capacities that emerge from different uses of the same brain architecture. This is in line with the past theses introduced under the terms of modularization (Karmiloff-Smith 1992), neuronal recycling (Dehaene and Cohen 2007; Peretz et al. 2015), neural reuse (Anderson 2010), or neural retuning (Matchin 2018).

## Conclusions and future directions

This paper investigated the continuum between nonhuman primates' cognitive capacities and human hierarchical structure building capacity, and integrated the findings of

comparative research into a working memory framework. Hierarchical structure building requires working memory to maintain and manipulate intermediate results based on stored long-term memory representations. In particular, I argued that the evolution of hierarchical structure building capacity for language and music can be investigated in terms of the action-related components such as goal, planning, control, and sensory-motor integration by focusing on the hierarchical control of goal-directed action implemented in the cortico-basal ganglia-thalamocortical (CBGT) circuits and hierarchical internal models realized by the dorsal pathways. The basic brain architecture crucial for hierarchical structure building in language and music is shared with non-human primates, but has elaborated in the course of hominin evolution to accomplish more complex hierarchical structure building in sequencing. This elaboration went hand in hand with the enhancement of working memory. Moreover, I proposed that one possible explanation for functional specialization of the brain architecture for language and music is a working memory resource optimization strategy. This line of comparative research from a bottom-up perspective makes it possible to study the evolution of hierarchical structure building capacity for language and music. However, this paper still leaves several questions open for future research.

For example, as reviewed above, hominin tool-making underwent gradual elaboration of action syntax and thus neuroanatomical change could directly relate to the hierarchical complexity of tool-making. On the other hand, to learn such a complex procedure, more complex social learning ability including imitation and teaching is required (Morgan et al. 2015; Laland 2017; Stout and Hecht 2017). It is also possible that the difference in the SLF III could reflect the difference in the social learning capacity (Hecht et al. 2013, 2015a). To learn how to use and make a tool by imitation, for example, one's own internal model should be linked to the state of the other, i.e., a simulation of the observed percept through one's own internal model is required (Wolpert et al. 2003; Stout and Hecht 2017). One possibility to disentangle this issue is comparative neuroanatomy of chimpanzee and bonobo (see, for example, Rilling et al. 2012): the former is a good tool user with less degree of prosociality, while the latter displays a high degree of prosociality and can socially learn tool-making, but rarely shows tool use in the wild (Wynn et al. 2011; and Gruber and Clay 2016 give good summaries of bonobo's tool-making ability). Therefore, research on action syntax also makes it possible to study why the basic brain architecture (or one particular subcomponent of it) crucial for hierarchical structure building was elaborated to yield more complex behavior.

Moreover, the bottom-up approach of this paper can also profit much from the progress made in the research on vocal learning across species (e.g., Petkov and Jarvis 2012; Lattenkamp and Vernes 2018; Jarvis 2019; Martins and Boeckx 2020). In particular, research on songbirds could contribute to revealing the brain mechanisms underlying chunking and rule-learning. In learning songs, juvenile songbirds break tutor songs into smaller units, produce those units as chunks to practice singing, and recombine those chunks to create their songs with large individual repertoires (Marler 2000; ten Cate and Okanoya 2012). Thus, chunks are units of manipulation in song production and perception as well as rule-learning. In non-vocal learning species, gibbons could be an additional model to investigate chunk-based vocal control as they show chunk-based organization in their song-like vocalization (Inoue et al. 2017).

Finally, the brain architecture discussed in the current paper points to a particular computational neurocognitive modeling strategy, namely hierarchical reinforcement learning, which unifies hierarchical control of goal-directed action and hierarchical internal models (e.g., Haruno et al. 2003; Frank and Badre 2012; Alexander and Brown 2018). This modeling strategy in combination with the evolutionary simulation of recursive combination as suggested by Toya and Hashimoto (2018) could provide us with a possibility to investigate why hierarchical structure building capacity has evolved in its full range of complexity in the hominin line. This allows language and music evolution research to move toward an integrated approach discussed elsewhere (Asano and Seifert 2018). Altogether, I hope the current paper made the evolution of hierarchical structure building capacity for language and music more tractable for comparative evolutionary research and offered possible future directions for interdisciplinary research.

**Acknowledgements** This work was supported by MEXT/JSPS Grant-in-Aid for Scientific Research on Innovative Areas #4903 (Evolinguistics) [grant number JP17H06379].

**Funding** Open Access funding enabled and organized by Projekt DEAL.

## Declarations

**Conflict of interest** None.

## References


Aboitiz F (2017) A brain for speech. Palgrave Macmillan

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

Aboitiz F, García RR, Bosman C, Brunetti E (2006) Cortical memory mechanisms and language origins. Brain Lang 98:40–56. https://doi.org/10.1016/j.bandl.2006.01.006

Aboitiz F, Aboitiz S, García RR (2010) The phonological loop. A key innovation in human evolution. Curr Anthropol 51:S55–S65. https://doi.org/10.1086/650525

Alexander WH, Brown JW (2018) Frontal cortex function as derived from hierarchical predictive coding. Sci Rep 8:1–11. https://doi.org/10.1038/s41598-018-21407-9

Alexander GE, Delong MR, Strick PL (1986) Parallel organization of functionally segregated circuits linking basal ganglia and cortex. Annu Rev Neurosci 9:357–381. https://doi.org/10.1146/annurev.ne.09.030186.002041

Ambrose SH (2001) Paleolithic technology and human evolution. Science 291:1748–1753. https://doi.org/10.1126/science.1059487

Anderson ML (2010) Neural reuse: a fundamental organizational principle of the brain. Behav Brain Sci 33:245–266. https://doi.org/10.1017/S0140525X10000853

Asano R (2019) Principled explanations in comparative biomusicology – toward a comparative cognitive biology of the human capacities for music and language. University of Cologne

Asano R, Boeckx C (2015) Syntax in language and music: what is the right level of comparison? Front Psychol 6:00942. https://doi.org/10.3389/fpsyg.2015.00942

Asano R, Seifert U (2018) Commentary: the evolution of musicality: what can be learned from language evolution research? Front Neurosci 12:640. https://doi.org/10.3389/fnins.2018.00640

Baddeley A (2010) Working memory. Curr Biol 20:R136–R140. https://doi.org/10.1016/j.cub.2009.12.014

Baddeley A (2012) Working memory: theories, models, and controversies. Annu Rev Psychol 63:1–29. https://doi.org/10.1146/annurev-psych-120710-100422

Badre D (2008) Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes. Trends Cogn Sci 12:193–200. https://doi.org/10.1016/j.tics.2008.02.004

Badre D, D'Esposito M (2009) Is the rostro-caudal axis of the frontal lobe hierarchical? Nat Rev Neurosci 10:659–669. https://doi.org/10.1038/nrn2667

Badre D, Desrochers TM (2019) Hierarchical cognitive control and the frontal lobes. In: D'Esposito M, Grafman J (eds) Handbook of clinical neurology, 3rd edn. Elsevier, pp 165–177

Badre D, Nee DE (2018) Frontal cortex and the hierarchical control of behavior. Trends Cogn Sci 22:170–188. https://doi.org/10.1016/j.tics.2017.11.005

Balezeau F, Wilson B, Gallardo G et al (2020) Primate auditory prototype in the evolution of the arcuate fasciculus. Nat Neurosci 23:611–614. https://doi.org/10.1038/s41593-020-0623-9

Balleine BW, Dezfouli A, Ito M, Doya K (2015) Hierarchical control of goal-directed action in the cortical–basal ganglia network. Curr Opin Behav Sci 5:1–7. https://doi.org/10.1016/j.cobeha.2015.06.001

Barton RA (2012) Embodied cognitive evolution and the cerebellum. Philos Trans R Soc B Biol Sci 367:2097–2107. https://doi.org/10.1098/rstb.2012.0112

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

Boeckx C (2013) Biolinguistics: forays into human cognitive biology. J Anthropol Sci 91:1–28. https://doi.org/10.4436/jass.91009

Boeckx C (2017) Language evolution. In: Kaas JH (ed) Evolution of nervous systems, 2nd edn. Elsevier, pp 325–339

Bornkessel-schlesewsky I, Schlesewsky M, Small SL, Rauschecker JP (2015) Neurobiological roots of language in primate audition: common computational properties. Trends Cogn Sci 19:1–9. https://doi.org/10.1016/j.tics.2014.12.008

Bornkessel-Schlesewsky I, Schlesewsky M, Small SL, Rauschecker JP (2015) Neurobiological roots of language in primate audition: common computational properties. Trends Cogn Sci 19:142–150. https://doi.org/10.1016/j.tics.2014.12.008

Bostan AC, Strick PL (2018) The basal ganglia and the cerebellum: nodes in an integrated network. Nat Rev Neurosci 19:338–350. https://doi.org/10.1038/s41583-018-0002-7

Bril B, Rein R, Nonaka T et al (2010) The role of expertise in tool use: skill differences in functional action adaptations to task constraints. J Exp Psychol Hum Percept Perform 36:825–839. https://doi.org/10.1037/a0018171

Brown S, Yuan Y, Belyk M (2020) Evolution of the speech-ready brain: the voice/jaw connection in the human motor cortex. J Comp Neurol. https://doi.org/10.1002/cne.24997

Bruner E (2004) Geometric morphometrics and paleoneurology: brain shape evolution in the genus Homo. J Hum Evol 47:279–303. https://doi.org/10.1016/j.jhevol.2004.03.009

Bruner E (2010) Morphological differences in the parietal lobes within the human genus. Curr Anthropol 51:S77–S88. https://doi.org/10.1086/650729

Buchsbaum BR, D'Esposito M (2019) A sensorimotor view of verbal working memory. Cortex 112:134–148. https://doi.org/10.1016/j.cortex.2018.11.010

Byrne RW, Russon AE (1998) Learning by imitation: a hierarchical approach. Behav Brain Sci 21:667–684. https://doi.org/10.1017/S0140525X98001745

Byrne RW, Sanz CM, Morgan DB (2013) Chimpanzees plan their tool use. In: Sanz CM, Call J, Boesch C (eds) Tool use in animals. Cambridge University Press, pp 48–64

Cannon JJ, Patel AD (2021) How beat perception co-opts motor neurophysiology. Trends Cogn Sci 25:137–150. https://doi.org/10.1016/j.tics.2020.11.002

Christiansen MH, Chater N (2016) The now-or-never bottleneck: a fundamental constraint on language. Behav Brain Sci 39:e62. https://doi.org/10.1017/S0140525X1500031X

Coolidge FL, Wynn T (2005) Working memory, its executive functions, and the emergence of modern thinking. Cambridge Archaeol J 15:5–26. https://doi.org/10.1017/S0959774305000016

Coolidge FL, Wynn T (2007) The working memory account of Neandertal cognition—how phonological storage capacity may be related to recursion and the pragmatics of modern speech. J Hum Evol 52:707–710. https://doi.org/10.1016/j.jhevol.2007.01.003

Cowan N (2010) The magical mystery four. Curr Dir Psychol Sci 19:51–57. https://doi.org/10.1177/0963721409359277

D'Mello AM, Gabrieli JDE, Nee DE (2020) Evidence for hierarchical cognitive control in the human cerebellum. Curr Biol 30:1–12. https://doi.org/10.1016/j.cub.2020.03.028

Dehaene S, Cohen L (2007) Cultural recycling of cortical maps. Neuron 56:384–398. https://doi.org/10.1016/j.neuron.2007.10.004

Fischer J, Hammerschmidt K (2020) Towards a new taxonomy of primate vocal production learning. Philos Trans R Soc B Biol Sci 375:20190045. https://doi.org/10.1098/rstb.2019.0045

Fitch WT (2006) The biology and evolution of music: a comparative perspective. Cognition 100:173–215. https://doi.org/10.1016/j.cognition.2005.11.009

Fitch WT (2011) The evolution of syntax: an exaptationist perspective. Front Evol Neurosci 3:9. https://doi.org/10.3389/fnevo.2011.00009
Fitch WT (2017) Empirical approaches to the study of language evolution. Psychon Bull Rev 24:3–33. https://doi.org/10.3758/s13423-017-1236-5
Fitch WT, Martins MD (2014) Hierarchical processing in music, language, and action: Lashley revisited. Ann N Y Acad Sci 1316:87–104. https://doi.org/10.1111/nyas.12406
Frank MJ, Badre D (2012) Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: computational analysis. Cereb Cortex 22:509–526. https://doi.org/10.1093/cercor/bhr114
Friederici AD (2019) Hierarchy processing in human neurobiology: how specific is it? Philos Trans R Soc B Biol Sci 375:20180391. https://doi.org/10.1098/rstb.2018.0391
Fritz J, Mishkin M, Saunders RC (2005) In search of an auditory engram. Proc Natl Acad Sci 102:9359–9364. https://doi.org/10.1073/pnas.0503998102
Ghazanfar AA, Takahashi DY (2014a) Facial expressions and the evolution of the speech rhythm. J Cogn Neurosci 26:1196–1207. https://doi.org/10.1162/jocn_a_00575
Ghazanfar AA, Takahashi DY (2014b) The evolution of speech: vision, rhythm, cooperation. Trends Cogn Sci 18:543–553. https://doi.org/10.1016/j.tics.2014.06.004
Gruber T, Clay Z (2016) A comparison between bonobos and chimpanzees: a review and update. Evol Anthropol News, Rev 25:239–252. https://doi.org/10.1002/evan.21501
Haber SN (2003) The primate basal ganglia: parallel and integrative networks. J Chem Neuroanat 26:317–330. https://doi.org/10.1016/j.jchemneu.2003.10.003
Haber SN (2016) Corticostriatal circuitry. Dialogues in Clinical Neuroscience
Hage SR, Nieder A (2016) Dual neural network model for the evolution of speech and language. Trends Neurosci 39:813–829. https://doi.org/10.1016/j.tins.2016.10.006
Haruno M, Wolpert DM, Kawato M (2003) Hierarchical MOSAIC for movement generation. Int Congr Ser 1250:575–590. https://doi.org/10.1016/S0531-5131(03)00190-0
Hauser MD, Yang C, Berwick RC et al (2014) The mystery of language evolution. Front Psychol 5:401. https://doi.org/10.3389/fpsyg.2014.00401
Hayashi M (2007) A new notation system of object manipulation in the nesting-cup task for chimpanzees and humans. Cortex 43:308–318. https://doi.org/10.1016/S0010-9452(08)70457-X
Hecht EE, Gutman DA, Preuss TM et al (2013) Process versus product in social learning: comparative diffusion tensor imaging of neural systems for action execution-observation matching in macaques, chimpanzees, and humans. Cereb Cortex 23:1014–1024. https://doi.org/10.1093/cercor/bhs097
Hecht EE, Gutman DA, Bradley BA et al (2015a) Virtual dissection and comparative connectivity of the superior longitudinal fasciculus in chimpanzees and humans. Neuroimage 108:124–137. https://doi.org/10.1016/j.neuroimage.2014.12.039
Hecht EE, Gutman DA, Khreisheh N et al (2015b) Acquisition of paleolithic toolmaking abilities involves structural remodeling to inferior frontoparietal regions. Brain Struct Funct 220:2315–2331. https://doi.org/10.1007/s00429-014-0789-6
Hickok G (2012) Computational neuroanatomy of speech production. Nat Rev Neurosci 13:135–145. https://doi.org/10.1038/nrn3158
Hickok G (2017) A cortical circuit for voluntary laryngeal control: implications for the evolution language. Psychon Bull Rev 24:56–63. https://doi.org/10.3758/s13423-016-1100-z
Hickok G, Poeppel D (2004) Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92:67–99. https://doi.org/10.1016/j.cognition.2003.10.011
Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nat Rev Neurosci 8:393–402. https://doi.org/10.1038/nrn2113
Holloway RL (2015) The evolution of the hominid brain. In: Henke W, Tattersall I (eds) Handbook of paleoanthropology. Springer, pp 1961–1987
Hopkins WD, Latzman RD, Mareno MC et al (2019) Heritability of gray matter structural covariation and tool use skills in chimpanzees (Pan troglodytes): a source-based morphometry and quantitative genetic analysis. Cereb Cortex 29:3702–3711. https://doi.org/10.1093/cercor/bhy250
Inoue Y, Sinun W, Yosida S, Okanoya K (2017) Combinatory rules and chunk structure in male Mueller's gibbon songs. Interact Stud Soc Behav Commun Biol Artif Syst 18:1–25. https://doi.org/10.1075/is.18.1.01ino
Ito M (2008) Control of mental activities by internal models in the cerebellum. Nat Rev Neurosci 9:304–313. https://doi.org/10.1038/nrn2332
Jackendoff R (2009) Parallels and nonparallels between language and music. Music Percept 26:195–204. https://doi.org/10.1525/mp.2009.26.3.195
Jarvis ED (2019) Evolution of vocal learning and spoken language. Science 366:50–54. https://doi.org/10.1126/science.aax0287
Jeon H-A (2014) Hierarchical processing in the prefrontal cortex in a variety of cognitive domains. Front Syst Neurosci 8:1–8. https://doi.org/10.3389/fnsys.2014.00223
Jeon H-A, Friederici AD (2015) Degree of automaticity and the prefrontal cortex. Trends Cogn Sci 19:244–250. https://doi.org/10.1016/j.tics.2015.03.003
Jürgens U (2002) Neural pathways underlying vocal control. Neurosci Biobehav Rev 26:235–258. https://doi.org/10.1016/S0149-7634(01)00068-9
Karmiloff-Smith A (1992) Beyond modularity. MIT Press
Karmiloff-Smith A (2013) Challenging the use of adult neuropsychological models for explaining neurodevelopmental disorders: developed versus developing brains. Q J Exp Psychol 66:1–14. https://doi.org/10.1080/17470218.2012.744424
Kljajevic V (2010) Is syntactic working memory language specific? Psihologija 43:85–101. https://doi.org/10.2298/PSI1001085K
Koda H, Kunieda T, Nishimura T (2018) From hand to mouth: monkeys require greater effort in motor preparation for voluntary control of vocalization than for manual actions. R Soc Open Sci 5:180879. https://doi.org/10.1098/rsos.180879
Koelsch S (2012) Brain and music. Wiley-Blackwell
Koelsch S, Vuust P, Friston K (2019) Predictive processes and the peculiar case of music. Trends Cogn Sci 23:63–77. https://doi.org/10.1016/j.tics.2018.10.006
Kotz SA, Schwartze M, Schmidt-Kassow M (2009) Non-motor basal ganglia functions: a review and proposal for a model of sensory predictability in auditory language perception. Cortex 45:982–990. https://doi.org/10.1016/j.cortex.2009.02.010
Kumar V, Croxson PL, Simonyan K (2016) Structural organization of the laryngeal motor cortical network and its implication for evolution of speech production. J Neurosci 36:4170–4181. https://doi.org/10.1523/JNEUROSCI.3914-15.2016
Kuypers HGJM (1958a) Corticobular connexions to the pons and lower brain-stem in man: an anatomical study. Brain 81:364–388. https://doi.org/10.1093/brain/81.3.364
Kuypers HGJM (1958b) Some projections from the peri-central cortex to the pons and lower brain stem in monkey and chimpanzee. J Comp Neurol 110:221–255. https://doi.org/10.1002/cne.901100205
Laland KN (2017) The origins of language in teaching. Psychon Bull Rev 24:225–231. https://doi.org/10.3758/s13423-016-1077-7
Lameira AR (2017) Bidding evidence for primate vocal learning and the cultural substrates for speech evolution. Neurosci Biobehav

Rev 83:429–439. https://doi.org/10.1016/j.neubiorev.2017.09.021
Lashley K (1951) The problem of serial order in behavior. In: Jeffress LA (ed) Cerebral mechanisms in behavior: the Hixon symposium. Wiley, pp 112–147
Lattenkamp EZ, Vernes SC (2018) Vocal learning: a language-relevant trait in need of a broad cross-species approach. Curr Opin Behav Sci 21:209–215. https://doi.org/10.1016/j.cobeha.2018.04.007
Lerdahl F, Jackendoff R (1983) A generative theory of tonal music. MIT Press
Lieberman P (2016) The evolution of language and thought. J Anthropol Sci 94:1–20. https://doi.org/10.4436/jass.94029
Loh KK, Petrides M, Hopkins WD et al (2017) Cognitive control of vocalizations in the primate ventrolateral-dorsomedial frontal (VLF-DMF) brain network. Neurosci Biobehav Rev 82:32–44. https://doi.org/10.1016/j.neubiorev.2016.12.001
MacNeilage PF (1998) The frame/content theory of evolution of speech production. Behav Brain Sci 21:499–511
Marler P (2000) Origins of music and speech: Insights from animals. In: Wallin N, Merker B, Brown S (eds) The origins of music. MIT Press, pp 31–48
Martins PT, Boeckx C (2020) Vocal learning: beyond the continuum. PLOS Biol 18:e3000672. https://doi.org/10.1371/journal.pbio.3000672
Matchin WG (2018) A neuronal retuning hypothesis of sentence-specificity in Broca's area. Psychon Bull Rev 25:1682–1694. https://doi.org/10.3758/s13423-017-1377-6
Matsuzawa T (1991) Nesting cups and metatools in chimpanzees. Behav Brain Sci 14:570–571
Matsuzawa T (1996) Chimpanzee intelligence in nature and in captivity: isomorphism of symbol use and tool use. In: McGrew WC, Marchant LF, Nishida T (eds.) Great ape societies. Cambridge University Press, pp 196–210
Merchant H, Honing H (2014) Are non-human primates capable of rhythmic entrainment? Evidence for the gradual audiomotor evolution hypothesis. Front Neurosci 7:274. https://doi.org/10.3389/fnins.2013.00274
Middleton FA, Strick PL (2000a) Basal ganglia and cerebellar loops: motor and cognitive circuits. Brain Res Rev 31:236–250. https://doi.org/10.1016/S0165-0173(99)00040-5
Middleton FA, Strick PL (2000b) Basal ganglia output and cognition: evidence from anatomical, behavioral, and clinical studies. Brain Cogn 42:183–200. https://doi.org/10.1006/brcg.1999.1099
Moore MW (2010) "Grammars of action" and stone flaking design space. In: Nowell A, Davidson I (eds) Stone tools and the evolution of human cognition. University Press of Colorado, pp 13–43
Morgan TJH, Uomini NT, Rendell LE et al (2015) Experimental evidence for the co-evolution of hominin tool-making teaching and language. Nat Commun 6:6029. https://doi.org/10.1038/ncomms7029
Neubauer S, Hublin J-J, Gunz P (2018) The evolution of modern human brain shape. Sci Adv 4:eaao5961. https://doi.org/10.1126/sciadv.aao5961
Nonaka T, Bril B, Rein R (2010) How do stone knappers predict and control the outcome of flaking? Implications for understanding early stone tool technology. J Hum Evol 59:155–167. https://doi.org/10.1016/j.jhevol.2010.04.006
Patel AD (2008) Music, language, and the brain. Oxford University Press
Patel AD (2012) Language, music, and the brain: a resource-sharing framework. In: Rebuschat P, Rohmeier M, Hawkins JA, Cross I (eds) Language and Music as Cognitive Systems. Oxford University Press, pp 204–223
Patel AD (2013) Sharing and Nonsharing of Brain Resources for Language and Music. In: Arbib MA (ed) Language, music, and the brain. MIT Press, pp 329–355
Patel AD, Iversen JR (2014) The evolutionary neuroscience of musical beat perception: the action simulation for auditory prediction (ASAP) hypothesis. Front Syst Neurosci 8:57. https://doi.org/10.3389/fnsys.2014.00057
Peretz I (2013) The biological foundations of music: insights from congenital amusia. In: Deutsch D (ed) The psychology of music, 3rd edn. Academic Press, pp 551–564
Peretz I, Vuvan D, Lagrois M-É, Armony JL (2015) Neural overlap in processing music and speech. Philos Trans R Soc Lond B Biol Sci 370:20140090. https://doi.org/10.1098/rstb.2014.0090
Petkov CI, Jarvis ED (2012) Birds, primates, and spoken language origins: behavioral phenotypes and neurobiological substrates. Front Evol Neurosci 4:12. https://doi.org/10.3389/fnevo.2012.00012
Petkov CI, ten Cate C (2020) Structured sequence learning: animal abilities, cognitive operations, and language evolution. Top Cogn Sci 12:828–842. https://doi.org/10.1111/tops.12444
Petkov CI, Wilson B (2012) On the pursuit of the brain network for proto-syntactic learning in non-human primates: conceptual issues and neurobiological hypotheses. Philos Trans R Soc B Biol Sci 367:2077–2088. https://doi.org/10.1098/rstb.2012.0073
Petrides M (2014) Neuroanatomy of language regions of the human brain. Academic Press
Petrides M, Cadoret G, Mackey S (2005) Orofacial somatomotor responses in the macaque monkey homologue of Broca's area. Nature 435:1235–1238. https://doi.org/10.1038/nature03628
Proksch S, Comstock DC, Médé B et al (2020) Motor and predictive processes in auditory beat and rhythm perception. Front Hum Neurosci. https://doi.org/10.3389/fnhum.2020.578546
Putt SS, Wijeakumar S, Franciscus RG, Spencer JP (2017) The functional brain networks that underlie early stone age tool manufacture. Nat Hum Behav 1:0102. https://doi.org/10.1038/s41562-017-0102
Quallo MM, Price CJ, Ueno K et al (2009) Gray and white matter changes associated with tool-use learning in macaque monkeys. Proc Natl Acad Sci 106:18379–18384. https://doi.org/10.1073/pnas.0909751106
Ramnani N (2006) The primate cortico-cerebellar system: anatomy and function. Nat Rev Neurosci 7:511–522. https://doi.org/10.1038/nrn1953
Rauschecker JP (2012) Ventral and dorsal streams in the evolution of speech and language. Front Evol Neurosci 4:7. https://doi.org/10.3389/fnevo.2012.00007
Rilling JK, Glasser MF, Preuss TM et al (2008) The evolution of the arcuate fasciculus revealed with comparative DTI. Nat Neurosci 11:426–428. https://doi.org/10.1038/nn2072
Rilling JK, Scholz J, Preuss TM et al (2012) Differences between chimpanzees and bonobos in neural systems supporting social cognition. Soc Cogn Affect Neurosci 7:369–379. https://doi.org/10.1093/scan/nsr017
Risueno-Segovia C, Hage SR (2020) Theta synchronization of phonatory and articulatory systems in marmoset monkey vocal production. Curr Biol 30:1–8. https://doi.org/10.1016/j.cub.2020.08.019
Scott BH, Mishkin M (2016) Auditory short-term memory in the primate auditory cortex. Brain Res 1640:264–277. https://doi.org/10.1016/j.brainres.2015.10.048
Scott BH, Mishkin M, Yin P (2012) Monkeys have a limited form of short-term memory in audition. Proc Natl Acad Sci 109:12237–12241. https://doi.org/10.1073/pnas.1209685109
Snyder B (2016) Memory for music, 2nd edn. Oxford University Press
Stout D (2010) The evolution of cognitive control. Top Cogn Sci 2:614–630. https://doi.org/10.1111/j.1756-8765.2009.01078.x

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

Stout D, Chaminade T (2007) The evolutionary neuroscience of tool making. Neuropsychologia 45:1091–1100. https://doi.org/10.1016/j.neuropsychologia.2006.09.014

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

Stout D, Hecht EE (2017) Evolutionary neuroscience of cumulative culture. Proc Natl Acad Sci 114:7861–7868. https://doi.org/10.1073/pnas.1620738114

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

Stout D, Passingham R, Frith C et al (2011) Technology, expertise and social cognition in human evolution. Eur J Neurosci 33:1328–1338. https://doi.org/10.1111/j.1460-9568.2011.07619.x

Stout D, Chaminade T, Thomik A et al (2018) Grammars of action in human behavior and evolution. biorXiv 3:281543. https://doi.org/10.1101/281543

ten Cate C, Okanoya K (2012) Revisiting the syntactic abilities of non-human animals: natural vocalizations and artificial grammar learning. Philos Trans R Soc B Biol Sci 367:1984–1994. https://doi.org/10.1098/rstb.2012.0055

Toya G, Hashimoto T (2018) Recursive combination has adaptability in diversifiability of production and material culture. Front Psychol 9:1512. https://doi.org/10.3389/fpsyg.2018.01512

Uhrig L, Dehaene S, Jarraya B (2014) A hierarchy of responses to auditory regularities in the macaque brain. J Neurosci 34:1127–1132. https://doi.org/10.1523/JNEUROSCI.3165-13.2014

Ullman MT (2006) Is Broca's area part of a basal ganglia thalamocortical circuit? Cortex 42:480–485. https://doi.org/10.1016/S0010-9452(08)70382-4

Wakita M (2020) Language evolution from a perspective of Broca's area. In: Masataka N (ed) The origins of language revisited. Springer Nature, pp 97–113

Weiss DJ, Chapman KM, Wark JD, Rosenbaum DA (2012) Motor planning in primates. Behav Brain Sci 35:244–244. https://doi.org/10.1017/S0140525X1100197X

Wilson B, Marslen-Wilson WD, Petkov CI (2017) Conserved sequence processing in primate frontal cortex. Trends Neurosci 40:72–82. https://doi.org/10.1016/j.tins.2016.11.004

Wilson B, Spierings M, Ravignani A et al (2020) Non-adjacent dependency learning in humans and other animals. Top Cogn Sci 12:843–858. https://doi.org/10.1111/tops.12381

Wolpert DM, Miall RC, Kawato M (1998) Internal models in the cerebellum. Trends Cogn Sci 2:338–347. https://doi.org/10.1016/S1364-6613(98)01221-2

Wolpert DM, Doya K, Kawato M (2003) A unifying computational framework for motor control and social interaction. Philos Trans R Soc London Ser B Biol Sci 358:593–602. https://doi.org/10.1098/rstb.2002.1238

Wynn T, Hernandez-Aguilar RA, Marchant LF, Mcgrew WC (2011) "An ape's view of the Oldowan" revisited. Evol Anthropol Issu News Rev 20:181–197. https://doi.org/10.1002/evan.20323