Title: The development of structured vocalizations in songbirds and humans: a comparative analysis

Dina Lipkind[1,2], Andreea Geambasu[3,4], Clara C. Levelt[3,4]

[1] Department of Psychology, Hunter College, The City University of New York, New York, NY, USA.

[2] Department of Biology, York College, The City University of New York, New York, NY, USA

[3] Centre for Linguistics, Leiden University, Leiden, The Netherlands.

[4] Leiden Institute for Brain and Cognition, Leiden University, Leiden, The Netherlands

Corresponding author: Dina Lipkind

Email**:** dina.lipkind@gmail.com

Mailing address: Department of Biology, School of Arts and Sciences,

York College, City University of New York,

94-20 Guy R. Brewer Blvd. Jamaica, NY, 11451

Keywords: Vocal learning, birdsong, human speech, vocal development, acoustic structure

Abstract

Humans and songbirds face a common challenge: acquiring the complex vocal repertoire of their social group. Although humans are thought to be unique in their ability to convey symbolic meaning through speech, speech and birdsong are comparable in their acoustic complexity and the mastery with which the vocalizations of adults are acquired by young individuals. In this review we focus on recent advances in the study of vocal development in humans and songbirds that shed new light on the emergence of distinct structural levels of vocal behavior, and point to new possible parallels between both groups.

1. Introduction

Vocal communication is common across a wide range of taxa, but the capacity to learn the acoustic structure of communicative vocalizations (vocal production learning) evolved in a rather small set of mammalian and avian species (reviewed in Petkov & Jarvis 2012). Among these, songbirds have been the oldest and most studied animal model of vocal learning. The similarities between the developmental learning of birdsong and human speech have been long noted (reviewed in Doupe & Kuhl 1999), strongly suggesting convergent evolution of learning mechanisms across these phylogenetically distant groups. Like speech, birdsongs are highly structured vocalizations that are culturally transmitted to young individuals by adult conspecifics, often during restricted developmental time windows. However, unlike human language, birdsongs do not seem to convey symbolic meaning, except perhaps in a very rudimentary way (Suzuki et al. 2016; Engesser et al. 2016). Therefore, both structurally and developmentally, birdsong and speech can be most fruitfully compared at the level of phonology, the linguistic level addressing sound structure, from phonetic features to intonational phrases (Doupe & Kuhl 1999; Mol et al. 2017; Yip 2013).

The last two decades of birdsong research have seen several advances in elucidating the developmental processes and neural mechanisms that mediate the learning of distinct levels of song structure, utilizing novel tools for the experimental manipulation and analysis of song development. The resulting findings, together with parallel progress in speech development studies, allow us to attempt a more detailed alignment between distinct developmental processes in birdsong and human speech than was previously possible.

Here we review recent insights from both fields on two key aspects of vocal development: the emergence and shaping of individual units of vocal production and the

acquisition of the ability to combine vocal units into diverse sequences, focusing on new possible parallels between birdsong and human speech.

2. Development of vocal production units – from analog to discrete performance

Like human speech, the songs of most songbird species are highly structured (see Table 1 for terminology): a typical song consists of syllables – sound bursts separated by brief silent intervals, which are sometimes composed of smaller elements – notes. Syllables are grouped into sequences termed phrases or motifs, which are in turn grouped to form song bouts (e.g., the zebra finch song in Fig. 1a). In both humans and songbirds, the components of mature vocal performance can be readily classified into discrete categories. Speech components – e.g., phonemes and syllables – consist of finite sets specified by a given language. Similarly, birdsong syllables, phrases, motifs, and whole songs often fall into a small number of acoustically distinct "types" which appear as clusters in distributions of acoustic parameters of individual renditions (Fig. 1a; Wohlgemuth et al. 2010; Derégnaucourt et al. 2005; Tchernichovski et al. 2004; Janney et al. 2016; Sasahara et al. 2015). However, these highly differentiated acoustic categories of adult performance emerge from early vocalizations that are graded, or analog, in nature – varying along continuous acoustic parameters rather than across discrete states (Fig. 1b). Vocal development in both songbirds and human infants thus shows a gradual emergence of discrete vocal categories from highly variable and unstructured performance. This transition appears to be a combination of universal (non-learned) developmental processes, and of processes that are shaped by sensory input from the vocalizations of adults.

----------Insert Table 1 about here----------

2.1 Emergence of vocal units in birdsong

Birdsong development, which has been extensively studied in zebra finches - the main model species for birdsong learning, begins with subsong – the immature singing of juveniles. Subsong is initially a graded and amorphous signal, with no observable vocal categories (Fig. 1b). The earliest observable structural regularity in subsong is the appearance of rhythmical performance of repeated syllable "prototypes" with relatively stereotyped durations, but variable acoustic structure. Thus, "coarse" temporal structuring of syllable durations precedes fine acoustic structuring. This process involves the development of a precisely coordinated activation of the avian vocal organ (syrinx) and respiration (Goller & Cooper 2004): in early subsong the relationship between breathing and vocalizing is irregular, but with the appearance of rhythmical syllable prototypes, syllables become fully synchronized with expirations and inter-syllabic gaps with inspirations (Veit et al. 2011; Aronov et al. 2011). Although temporal structuring often increases abruptly following exposure of naïve juveniles to adult song (Tchernichovski et al. 2001), it occurs also in juveniles that are isolated from male song (Mendez et al. 2010), and therefore is likely to constitute a largely pre-programmed developmental trend of increase in motor control and coordination.

In contrast with coarse temporal structuring, the development of the fine acoustic structure of syllable prototypes is strongly influenced by external auditory input, i.e., the singing of an adult "tutor" (usually the juvenile's father), and is marked by a gradual increase in acoustic similarity to the tutor's song (Tchernichovski et al. 2001). In parallel, the acoustic variability of syllable performance gradually decreases (Ravbar et al. 2012; Derégnaucourt et al. 2005). Together these two developmental trends constitute the differentiation of amorphous syllable prototypes into distinct syllable "types" (Fig. 1 a-b). Interestingly, a

single syllable prototype can "duplicate", generating multiple syllable types (reminiscent of the division and differentiation of cells). Consecutive renditions of the same prototype, acoustically indistinguishable early in development, become increasingly dissimilar and end up resembling different syllables in the target song (Fig. 1c; Tchernichovski et al. 2001; Liu et al. 2004). The differentiation of syllable performance is mirrored by a gradual differentiation of neural activity in a premotor cortical area specialized for song (Okubo et al. 2015).

2.2 The role of motor variability in birdsong learning

The gradual reduction in performance variability, which accompanies the fine structuring of syllables, has been the subject of a series of behavioral and neural studies, providing new insights on the causes and function of performance variability in developmental vocal learning. Traditionally viewed as stemming from poor motor control in young and inexperienced performers, vocal variability in songbirds was shown to be actively generated and regulated via specialized neural circuits. A component of a basal ganglia-cortical circuit specialized for song learning, the anterior forebrain pathway, is necessary for generating variable performance (Aronov et al. 2008): its inactivation in juvenile zebra finches results in a transition to stereotyped performance, in effect "freezing" song development (Olveczky et al. 2005). Moreover, the acoustic variability of developing syllables is regulated such that it is high when performance is off target and becomes lower with improved imitation. This regulation of variability occurs at multiple time scales, ranging from milliseconds to weeks. For example, within a given song motif, the performance of syllables that are still in the process of being learned is considerably more variable than that of syllables that have reached their learning target (Fig. 1d; Ravbar et al. 2012),

demonstrating birds' ability to control vocal variability on a moment-to-moment basis. This presumably allows a bird to work on specific parts of its song without destabilizing the performance of well-learned parts. Over diurnal cycles, vocal variability increases after night sleep, and decreases during the day, and the magnitude of these daily oscillations was shown to be correlated with learning success (Derégnaucourt et al. 2005). Finally, performance variability of individual syllables gradually decreases as they become more similar to the leaning target, a process that can take from several days to several weeks (Fig. 1d; Ravbar et al., 2012), and is mediated by the development of an inhibitory network which blocks auditory input from affecting the premotor control of song (Vallentin et al., 2016). Taken together, these findings demonstrate that vocal variability is not merely an obstacle that learners need to overcome. Instead, variability serves as a tool for motor exploration that facilitates the efficient learning of a complex vocal repertoire.

----------Insert Figure 1 about here----------

2.3 Emergence of vocal units in human speech

Similarly to birdsong development, speech development proceeds from highly amorphous and unstructured early vocalizations to the structured and relatively stereotyped performance of babbling and early words. In the earliest Phonation Stage, at 0-2 months (Oller 1980), infants begin to produce amorphous signals termed Quasi-Resonant-Nuclei: vowel-like and consonant-like sounds produced with (nearly) closed mouths. As in the subsong of songbirds, the first development, occurring at 2-3 months, is the appearance of temporal structuring. In human infants this is achieved by interruptions of the breathing cycle, resulting from erratic contact between the tongue dorsum and the palate, which give rise to typical "coo" sounds. The Cooing (or GOO) Stage, is followed by a period

characterized by a great variability in sounds and sound qualities, the Expansion Stage (at 4-6 months), in which infants experiment with repetitive productions of now fully resonant nuclei (vowels), squeals, growls and labio-lingual trills called "raspberries". This period of self-monitored physical exploration leads to the formation of primitive sound categories (Oller & Griebel 2008) – consonants (C) and vowels (V).

The next development, around 6-8 months, starts out with a rhythmic opening and closing of the jaw, initially a general motor stereotypy that coincides with increased rhythmic arm movements (Locke et al. 1995; Ejiri 1998). According to the Frame-Content model (MacNeilage & Davis, 1990), coordination of these oscillating jaw movements (creating a *frame*) with varying tongue positions (creating *content*) results in the performance of a core structural unit across languages - the Consonant Vowel (CV) syllable. This constitutes the start of the Canonical Babbling stage (Stark 1980; Smith et al. 1989; Oller 1980; Geambasu, Scheel & Levelt 2016). Together, the cooing, expansion, and canonical babbling stages constitute a largely universal process of acquiring the ability for coordinated activation of breathing and lower and upper vocal tract articulators, a process that is analogous to juvenile songbirds' mastering of the ability to perform rhythmical proto-syllables.

The new motor capacity of canonical babbling brings the infant's vocal production closer to sounding like language, and this, in turn, affects the quality of the input from the infant's "tutors", providing the infant with more directed, language-specific acoustic targets and feedback (Goldstein et al. 2003). As a result, the relative frequencies and acoustic properties of syllables produced by infants gradually shift towards an increased resemblance with the ambient language (Sagart & Durand 1984; De Boysson-Bardies & Vihman 1991).

While the influence of the ambient language can thus already be discerned in babbling, it is more amenable to investigation in early word production, since it is then

possible to know exactly which sounds the child is targeting. Motor patterns in babbling and early word-productions initially overlap: infants tend to use well-established motor patterns from babbling to produce words with similar sound characteristics. A characteristic example (Waterson,1971) is a child producing the same sequence [baebu:] for multiple bisyllabic target words, *Patrick, Bobby, birdie, bucket* and *button*, all of which start with a labial plosive and contain a medial plosive. Over time, the child's productions of these different words become increasingly dissimilar and more target-like (e.g. [baebu:] > [bʌtɪk] > [baetɪk] > [paetɪk], for *Patrick*). Note that the phonemes /p/ and /b/, are initially realized in a non-contrasting way [b]. However, acoustic analyses have identified *covert* contrasts in children's realizations of phonemes; phonemes like /p/ and /b/ are differentiated in production by the developing speaker, but in ways that are imperceptible to the adult ear (Scobbie et al. 2000; McAllister-Byun et al. 2016). This gradual developmental process is reminiscent of the duplication and differentiation of syllable types seen in zebra finches.

2.4 The role of variability in speech development

Performance variability may be a necessary tool for the sensorimotor learning of the structural units of speech, as in birdsong. Vocal variability in infants has been measured in longitudinal recordings (Buder et al. 2003), but its function in learning the acoustic structure of speech sounds has not been specifically tested yet. Theoretical models of speech development, like the DIVA model (Guenther 1994, Guenther & Vladusich 2012) predict a role for performance variability in motor exploration. This model assumes that learning is driven by the initial mismatch between newly-acquired speech targets and the infant's production attempts. Variable performance during early development provides infants with auditory, motor, and somatosensory feedback, which is used to tune synaptic projections

between sensory and motor representations of speech sounds in the infant's brain. Oller and Griebel (2008) propose that there is a universal sequence of vocal events in human infants, starting with spontaneous production, which is subsequently elaborated through systematic vocal exploration of variations, leading to the formation of (primitive) categories. This cycle of production, exploration, and categorization is thought to apply to every new vocal domain and signal.

2.5 Summary: emergence of vocal production units

Both humans and songbirds progress from an early stage in which gross temporal structuring is achieved through increased coordination between muscles controlling the respiratory and vocal organs, and (in case of humans) the upper vocal tract articulators, resulting in the performance of basic vocal units that are further shaped by feedback from exposure to ambient song or speech. The process of developing a coordinated activation of the different muscle systems involved in vocal behavior is more complex in infants compared to songbirds, involving not only an early stage of coordinating breathing with phonation, but also a later stage (canonical babbling) of adding the coordinated performance of supra-glottal articulators. The role of the upper vocal tract is less dominant in birdsong production, in which sound is mostly structured by the syrinx. However, beak movements and upper vocal tract position were also found to contribute to sound structuring in adult zebra finches (Goller et al. 2004; Ohms et al. 2010), and parakeets (Ohms et al. 2012), raising the question of how and when this articulatory component emerges in the course of song development. Vocal variability may play an important role as a tuning mechanism in both songbirds and humans, by enabling young learners to explore diverse motor states and select those that lead to an

increased match with their target. This idea is currently supported by experimental findings in songbirds, and theoretical work on human vocal development.

3. Development of vocal combinatorial sequences

In parallel with learning the acoustic structure of individual vocal elements, learners need to obtain the correct sequencing of elements. The immense richness of human speech relies critically on vocal combinatorial ability - the ability to reuse a given set of structural units to generate diverse sequences. Similarly, many songbird species (e.g., starlings) are capable of generating variable sequences by reusing the same elements in different sequential contexts, and even zebra finches, whose natural song consists of a fixed syllable sequence, can be experimentally induced to rearrange learned syllables in a new order (Lipkind et al. 2013). Consequently, both humans and songbirds must possess dedicated plasticity mechanisms at the sequencing level. Research on such sequencing-specific aspects of vocal learning is still in its beginning, particularly in songbirds, but some insights are evident so far. One is that element pairs, or bigrams, play a dominant role in the learning of vocal sequences, both in constructing perceptual "templates" that guide vocal imitation, and as a constraint on production learning.

3.1 Development of vocal sequencing in songbirds

The sensory representations that shape the development of vocal sequences were studied in white-crowned sparrows by manipulating the sensory input available to birds (Rose et al. 2004; Plamondon et al. 2010). Surprisingly, juveniles that were reared in acoustic isolation and exposed to recordings of song phrase pairs (AB, BC, CD and DE), developed

species-typical multi-phrase song sequences (ABCDE). Thus, juveniles could concatenate auditory representations of phrase pairs into a single auditory template of an entire song. Exposure to reversed-order pairs (BA, CB, DC and ED) produced the reversed-order song EDCBA, but hearing single song phrases failed to elicit normal song sequences. These findings indicate that phrase bigrams contain necessary and sufficient information for guiding song sequence learning. Further evidence from Bengalese finches showed that not only the identities of element pairs, but also the frequencies of their performance are represented as a learning target. Bengalese finch songs contain points of variable syllable transitions (e.g., where syllable A can be followed either by syllable B or syllable C; Okanoya, 2004). Birds were trained to adjust the relative frequencies of alternative transitions to escape an aversive stimulus (a burst of loud noise) contingent on a specific transition. When training stopped, the transition frequencies spontaneously returned to their baseline values (Warren et al. 2012), pointing to the existence of a sensory representation of a "bigram syntax" that is actively maintained as a learning target.

On the motor production side, the ability to combine vocal units into sequences during vocal development was studied in two songbird species: zebra finches, which were experimentally induced to change syllable order in a learned song (ABC) to match a new target (ACB), and Bengalese finches whose mature songs naturally consist of variable syllable sequencing (Lipkind et al. 2013). In both species, new syllable sequencing was acquired slowly and laboriously, in a series of discrete steps, at which new pairwise transitions were added to the vocal repertoire one by one. This occurred even though birds were already proficient in performing the syllables themselves, pointing to the existence of a distinct mechanism for learning the sequential order of existing vocal units. Importantly, the slow acquisition of syllable transitions was not limited to syllables with specific acoustic properties, but was rather a general effect, suggesting that it is not related to difficulties in

transitioning between particular vocal gestures. What sort of mechanism could explain such general constraints on the development of vocal sequencing? A possible scenario, which still awaits experimental testing, is that vocal combinatorial ability in young learners is constrained by the slow development of a neural network connecting syllable representations to each other.

3.2 Development of vocal sequencing in infants

Remarkably, a similar stepwise process of acquiring combinatorial ability has been observed in the development of infant canonical babbling (using longitudinal data of English-acquiring infants (Lipkind et al. 2013). Infants appear to be constrained in incorporating newly learned CV syllables, into babbling utterances; initially, new syllables are performed predominantly in repetitive sequences (e.g. ga ga…), and only gradually begin to appear in variegated sequences (e.g. ga du ge…). This may indicate that, like songbirds, infants are initially limited in their ability to make transitions between different CV syllables. Another (not incompatible) possibility is that auditory feedback from repetitive syllable production may help strengthen connections between cortical areas that are activated by syllable production and perception, building strong motor memories and stable sensorimotor representations of syllables (Fagan 2015).

In infants, there is also evidence for constraints on sequencing that are specific to transitions between particular articulatory gestures. For example, similar invariant transitions have been observed in babbling sequences from different languages (MacNeilage et al., 2000; Oohashi, Watanabe & Taga, 2013), such as transitions between anterior-articulated to posterior-articulated consonants, but not between posterior-articulated to anterior-articulated consonants (Fig. 2) in CVC(V) sequences (e.g., Labial-Coronal [pata] but not Coronal-

Labial *[tapa], Labial-Dorsal [paka] but not Dorsal-Labial *[kapa] and Coronal-Dorsal [taka], but not Dorsal-Coronal *[kata]).

----------Insert Figure 2 about here----------

These invariant orders can remain in place for quite some time, and can even be transferred to first words (Ingram 1974; Fikkert & Levelt 2008). Posterior-to-anterior consonantal transitions within words usually take a long time to appear (around the age of 24 months), and developing speakers initially either simply avoid target words containing such sequences (Schwartz & Leonard 1982), or modify their production to include preferred sequences (examples from Dutch child language are shown in Table 2). Characteristically, this is done by changing the order of consonants (metathesis), or by a child-specific process called Consonant Harmony (see Levelt 2011, for an overview), resulting in a sequence of consonants with the same articulatory gesture, e.g. Labial-Labial.

----------Insert Table 2 about here----------

3.3 Summary: development of vocal sequencing

Both songbirds and humans show a gradual and highly constrained development of the ability to combine basic vocal units into sequences, which points to the possibility of convergent underlying neural mechanisms. Acquiring combinatorial ability at the level of CV syllables is obviously only a small component of the complex vocal sequencing abilities of humans. In songbirds, it remains an open question whether processes underlying the learning of syllable bigrams are sufficient to fully explain song sequencing, or whether distinct processes are involved in the learning of higher order sequences, such as the transitions between consecutive motifs (Hyland Bruno & Tchernichovski 2017).

4. The combined challenge of learning structural units and their sequencing

Despite evidence for distinct processes underlying the learning of vocal units and their sequential ordering, it is important to keep in mind that the distinction between units and their sequencing is not obvious in either perception or production. This is because the input that infants and juvenile songbirds receive from their caregivers, as well as their own vocal output, do not consist of units performed in isolation but of *fixed sequences* of units – words in humans and song motifs or phrases in songbirds.

In human language, the individual sounds of a word have to appear in a fixed order to provide access to its meaning. For example, in order to produce the word with the meaning "snow", the word's individual sounds [s], [n], and [o] have to appear in the order [$s_1n_2o_3$]. Any alternative order is considered incorrect by both listeners and speakers, because it does not match the order of the sounds as stored with the meaning "snow" in the mental lexicon. Birdsong motifs and phrases are clearly not words in the sense of meaningful units, but they resemble words in having a strict sequential order of sub-units. Moreover, in humans as well as songbirds the process of learning the structure of higher-order units such as words or song motifs, and of their composing sub-units (such as phonetic segments or song syllables) overlap in time. The tight coupling and the lack of obvious distinction between units and their sequencing has challenging implications, posing a difficult “choice” for learners between holistic and segmented representations of vocal sequences. For instance, a learner can employ a holistic strategy of treating the sound structure of the word "snow" as a single indivisible target, or extract a set of smaller targets ([s], [n], and [o]), which can be used to construct multiple words (e.g., "snow" and “nose”). Below we describe recent clues on how songbird and human learners deal with this dilemma.

4.1 Holistic versus segmented strategies for learning vocal sequences

Consider a young zebra finch performing strings of unformed proto-syllables ($P_1P_2P_3\ldots$), and attempting to learn a tutor song motif (ABC). The "pupil" must select a trajectory of vocal adjustments that would transform its own performance into the target. For example, the pupil can simply assign its own syllables to target syllables according to temporal order ($P_1 \rightarrow A$; $P_2 \rightarrow B$; $P_3 \rightarrow C$). However, if P_1 happens to be structurally more similar to C than to A, it might be preferable to assign C to it as a target, and then rearrange syllable order accordingly. The problem is that there are a vast number of possible combinations of structural and sequential adjustments that can transform one sequence into another. Consequently, selecting an optimal (or even a reasonably good) combination is a computationally intractable problem (Goldstein et al. 2006).

A recent study showed that, in mid-development, zebra finches obviate this problem by adopting a non-optimal strategy (Lipkind et al. 2017): they match every syllable in their own song to the most acoustically similar target syllable, completely disregarding sequential similarity, and then rearrange syllable order to correct sequence errors. For example, juveniles trained to learn the song ABC and then introduced to a new target song AC^+B (where C^+ is a slightly pitch-shifted version of C), first adjust syllable C to match the acoustically closest target C^+, despite its being at a different sequential position ($ABC \rightarrow ABC^+$). This results in a sequencing error, which birds then correct by rearranging syllable order ($ABC^+ \rightarrow AC^+B$). This strategy minimizes structural adjustments at the "price" of incurring increased sequencing costs. Interestingly, at earlier developmental stages, the opposite strategy of "whole motif" learning is observed (Liu et al. 2004; Okubo et al. 2015), in which a sequence of proto-syllables develops into an appropriately ordered target motif

without any sequential adjustments. Thus, zebra finches may switch from holistic matching strategies early in development to segmented matching strategies later on. Such non-optimal strategies (which minimize either structural or sequential changes) may have evolved to make the learning of vocal sequences computationally manageable.

A similar question has been a subject of several studies on human vocal development. During early speech production, words are thought to have holistic, rather than segmental, representations (Waterson, 1971; Ferguson & Farewell, 1975; Levelt 1994). Word templates with invariant sound sequences seem to be used; the developing speaker either selects target words from the ambient language that fit the template, or applies changes to make the word form fit the template. It is thought that only later do words become segmented into smaller units that can be handled independently (reused) in production. Vowels become independent from the template first, followed by word-initial consonants (Levelt 1994, Fikkert and Levelt 2008). The transition from word-like units to segmental units is thought to be determined by memory constraints, when the number of holistic word-representations reaches a critical mass, suggested to lie either between 50-100 words (Vihman & Velleman 1989) or 150-200 words (Sosa & Stoel-Gammon, 2006), and enforces a lexical reorganization (Macken 1979). This hypothesis still awaits rigorous testing, and in this context it is interesting to consider that a clearly-segmented learning strategy evolved in zebra finches, who learn just a single song. Thus, it is possible that segmented representations of fixed vocal sequences evolved as an adaptation reducing the computational complexity of vocal learning, maybe even prior to serving as memory-efficient representations of a very large vocal repertoire.

5. Conclusion

We have attempted to highlight similarities in the development of vocal units and their sequencing across humans and songbirds. Both start with spontaneous, amorphous productions, in the early subsong and phonation stages respectively, followed by coarse, and then fine, structuring of vocal building blocks. On the basis of the structural properties of early vocalizations – or rather the lack thereof – comparing the subsong stage in songbirds to the phonation stage in humans (as in Soha & Peters 2015) seems more fitting than the common comparison of subsong to infant babbling (e.g., Gobes & Bolhuis 2007; Goldstein et al. 2003; Mol et al. 2017). Variability plays an important role in learning the fine acoustic structure of individual sounds in birdsong and possibly also in speech. The combination of behavioral and neural studies in songbirds, and predictive models for speech development may inspire future research in the two fields. A transition from repetitive to diverse performance of vocal units may be central to the learning of both their structure and sequencing across species: repetition could function as a mechanism for forming stable and distinct sound-motor representations, while the capacity to transition between distinct sounds develops gradually with a stepwise addition of pairwise transitions to the vocal repertoire. Finally, humans and songbirds face similar challenges in the parallel learning of fixed vocal sequences such as words or song motifs, and the units they are composed of: both may share a developmental transition from holistic to segmented strategies for learning the fixed sequences in their vocal repertoire. We argue that this justifies a reappraisal of the idea that words and song motifs are not comparable (Yip 2013), which opens up a new and exciting prospect for comparative research.

References:

Aronov, D., Veit, L., Goldberg, J. H., & Fee, M. S. (2011). Two distinct modes of forebrain circuit dynamics underlie temporal patterning in the vocalizations of young songbirds. *The Journal of Neuroscience*, *31* (45), 16353–68.

Aronov, D., Andalman, A.S., & Fee, M.S. (2008). A Specialized Forebrain Circuit for Vocal Babbling in the Juvenile Songbird. *Science*, *320* (5876), 630–634.

Buder, E., Oller, D., & Magoon, J. (2003). Vocal intensity in the development of infant protophones. In: Solé, M., Recasans, D. & Romero, J. (Eds.), *Proceedings of the XVth International Congress of Phonetic Sciences*, pp. 2015-2018.

De Boysson-Bardies, B. & Vihman, M.M. (1991). Adaptation to language: Evidence from babbling and first words in four languages. *Language*, *67* (2), 297–319.

Derégnaucourt, S Mitra, P.P., Feher, O., Pytte, C. & Tchernichovski, O. (2005). How sleep affects the developmental learning of bird song. *Nature*, *433* (7027), 710-716.

Doupe, A. J. & Kuhl, P. K. (1999). Birdsong and human speech: Common Themes and Mechanisms. *Annual Review of Neuroscience*, *22* (1), 567–631.

Ejiri, K. (1998). Relationship between rhythmic behavior and canonical babbling in infant vocal development. *Phonetica*, *55*, 226-237.

Engesser, S., Ridley, A.R. & Townsend, S.W. (2016). Meaningful call combinations and compositional processing in the southern pied babbler. *Proceedings of the National Academy of Sciences*, *113* (21), 5976–5981.

Fagan, M.K. (2015). Why repetition? Repetitive babbling, auditory feedback, and cochlear implantation. *Journal of Experimental Child Psychology*, *137*, 125–136.

Ferguson, C. & Farwell, C. (1975). Words and Sounds in Early Language Acquisition. *Language*, *51* (2), 419–439.

Fikkert, P. & Levelt C. (2008). How does Place fall into place. The lexicon and emergent

constraints in children's developing grammars. In: Dresher, E. & K. Rice (Eds.) *Contrast in Phonology*. Berlin: Mouton, pp. 231-270.

Geambaşu, A., Scheel, M. & Levelt, C. (2016). Cross-linguistics patterns in infant babbling. In: Scott, D. & Waughtal, D. (Eds.), *Proceedings of the 40th Boston University Conference of Language Development*, Somerville, MA: Cascadilla Press, pp. 155-168.

Gobes, S.M.H. & Bolhuis, J.J. (2007). Birdsong Memory: A Neural Dissociation between Song Recognition and Production. *Current Biology*, *17* (9), 789–793.

Goldstein, A., Kolman, P. & Zheng, J. (2006). Minimum common string partition problem: Hardness and approximations. *Electronic Journal of Combinatorics*, *12* (1 R), 1–18.

Goldstein, M.H., King, A.P. & West, M.J. (2003). Social interaction shapes babbling: testing parallels between birdsong and speech. *Proceedings of the National Academy of Sciences of the United States of America*, *100* (13), 8030–5.

Goller, F. & Cooper, B.G. (2004). Peripheral motor dynamics of song production in the zebra finch. *Annals of the New York Academy of Sciences*, *1016*, 130–152.

Goller, F., Mallinckrodt, M. J. & Torti, S. D. (2004). Beak gape dynamics, during song in the zebra finch. *Journal of Neurobiology*, *59* (3), 289–303.

Guenther, F. (1994). A neural network model of speech acquisition and motor equivalent speech production. *Biological Cybernetics*, *72*, 43–53

Guenther, F. & Vladusich, T. (2012). A neural theory of speech acquisition and production. *Journal of Neurolinguistics*, *25* (5), 408-422.

Hyland Bruno, J. & Tchernichovski, O. (2017). Regularities in zebra finch song beyond the repeated motif. *Behavioural Processes*, (October), pp.1–7.

Ingram, D. (1974). Phonological rules in young children. *Journal of Child Language*, *1* (1), 49–64.

Janney, E., Taylor, H., Scharff, C., Rothenberg, D., Parra, L. C., & Tchernichovski, O.

(2016). Temporal regularity increases with repertoire complexity in the Australian pied butcherbird's song. *Royal Society Open Science*, *3* (9), 160357.

Levelt, C. (1994). *On the acquisition of Place.* PhD Dissertation, Leiden University. The Hague: Holland Academic Graphics.

Levelt, C. (2011). Consonant Harmony in child language. In: M. van Oostendorp, C. Ewen & K. Rice (Eds.). *Companion to Phonology*. Boston MA: Blackwell, 1691-1716.

Lipkind, D., Marcus, G. F., Bemis, D. K., Sasahara, K., Jacoby, N., Takahasi, M., Suzuki, K., Feher, O., Ravbar, P., Okanoya, K., & Tchernichovski, O. (2013). Stepwise acquisition of vocal combinatorial capacity in songbirds and human infants. *Nature*, *498* (7452):104-8.

Lipkind, D., Zai, A. T., Hanuschkin, A., Marcus, G. F., Tchernichovski, O., & Hahnloser, R. H. R. (2017). Songbirds work around computational complexity by learning song vocabulary independently of sequence. *Nature Communications*, *8* (1):1247

Liu, W.-C., Gardner, T. J. & Nottebohm, F. (2004). Juvenile zebra finches can use multiple strategies to learn the same song. *Proceedings of the National Academy of Sciences of the United States of America*, *101* (52), 18177–18182.

Locke, J., Bekken, K., McMinn-Larson, L. & Wein, D. (1995). Emergent control of manual and vocal-motor activity in relation to the development of speech. *Brain and Language 51*, 498–508.

Macken, M. (1979). Developmental reorganization of phonology: a hierarchy of basic units of acquisition. *Lingua 49*, 11–49.

MacNeilage, P. & Davis, B. (1990). Motor explanations of babbling and early speech patterns. In M. Jeannerod (Ed.) *Attention and performance XIII: motor representation and control*, Hillsdale, NJ: Lawrence Erlbaum, 567–582.

MacNeilage, P., Davis, B., Kinney, A., & Matyear, C. (2000). The motor core of speech: a

comparison of serial organization patterns in infants and languages. *Child Development*, *71* (1), 153–163.

McAllister Byun, T., Buchwald, A. & Mizoguchi, A. (2016). Covert contrast in velar fronting: an acoustic and ultrasound study. *Clinical Linguistics & Phonetics 30* (3-5), 249-276.

Mendez, J.M., Dall'Asén, A. G., Cooper, B. G., & Goller, F. (2010). Acquisition of an Acoustic Template Leads to Refinement of Song Motor Gestures. *Journal of Neurophysiology*, *104* (2), 984–993.

Mol, C., Chen, A., Kager, R. W. J., & Ter Haar, S. M. (2017). Prosody in birdsong: A review and perspective. *Neuroscience and Biobehavioral Reviews*, *81*, 167–180.

Ohms, V.R., Snelderwaard, P. Ch., Ten Cate, C., & Beckers, G. J. (2010). Vocal tract articulation in zebra finches. *PLoS One*, *30;5* (7):e11923.

Ohms, V. R., Beckers, G. J., ten Cate, C., & Suthers, R. A. (2012). Vocal tract articulation revisited: the case of the monk parakeet. *Journal of Experimental Biology*, *215* (Pt 1), 85-92.

Okanoya, K. (2004). The Bengalese finch: A window on the behavioral neurobiology of birdsong syntax. *Annals of the New York Academy of Sciences*, *1016*, 724–735.

Okubo, T. S., Mackevicius, E. L., Payne, H. L., Lynch, G. F. & Fee, M. S. (2015). Growth and splitting of neural sequences in songbird vocal development. *Nature*, *528* (7582), 352–357.

Oller, D.K. (1980). The emergence of the sounds of speech in infancy. In G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson, eds. *Child phonology 1 Production*. Academic Press, pp. 93–112.

Oller, D. K. & Griebel, U. (2008). Contextual flexibility in infant vocal development and the earliest steps in the evolution of language. In: Griebel, U. & Oller, D. (Eds.), *Evolution*

of communicative flexibility: Complexity, creativity and adaptability in human and animal communication, Cambridge, MA: The MIT Press, pp.141–168.

Olveczky, B. P., Andalman, A. S. & Fee, M. S. (2005). Vocal experimentation in the juvenile songbird requires a basal ganglia circuit. *PLoS biology*, *3* (5), e153.

Oohashi, H., Watanabe, H. & Taga, G. (2013). Development of a Serial Order in Speech Constrained by Articulatory Coordination. *PLoS One 8* (11): e78600.

Petkov, C. I. & Jarvis, E. D. (2012). Birds, primates, and spoken language origins: Behavioral phenotypes and neurobiological substrates. *Frontiers in Evolutionary Neuroscience*, *4*, 1–24.

Plamondon, S. L., Rose, G. J. & Goller, F. (2010). Roles of Syntax Information in Directing Song Development in White-Crowned Sparrows (Zonotrichia leucophrys). *J Comp Psychol.*, *124* (2), 117–132.

Ravbar, P., Lipkind, D., Parra, L. C. & Tchernichovski, O. 2012. Vocal exploration is locally regulated during song learning. *Journal of Neuroscience*, *32* (10), 3422-32.

Rose, G. J. et al. (2004). Species-typical songs in white-crowned sparrows tutored with only phrase pairs. *Nature*, *432* (7018), 753–8.

Sagart, L. & Durand, C. (1984). Discernible differences in the babbling of infants according to target language. *Journal of Child Language*, *11* (1), 1–15.

Sasahara, K. Tchernichovski, O., Takahasi, M., Suzuki, K. & Okanoya, K. (2015). A rhythm landscape approach to the developmental dynamics of birdsong. *Journal of The Royal Society Interface*, *12* (112), 20150802.

Schwartz, R. G., & Leonard, L. B. (1982). Do children pick and choose? An examination of phonological selection and avoidance in early lexical acquisition. *Journal of Child Language 9* (2), 319-336.

Scobbie, J., Gibbon, F., Hardcastle, W., & Fletcher, P. (2000). Covert contrast as a stage in

the acquisition of phonetics and phonology. In: M. B. Broe & J. B. Pierrehumbert (Eds.), *Papers in laboratory phonology V: Acquisition and the lexicon*. Cambridge: Cambridge University Press, pp. 194-207.

Smith, B. L., Brown-Sweeney, S. & Stoel-Gammon, C. (1989). A quantitative analysis of reduplicated and variegated babbling. *First Language*, *9*, 175–190.

Soha, J. A. & Peters, S. (2015). Vocal Learning in Songbirds and Humans: A Retrospective in Honor of Peter Marler. *Ethology*, *121* (10), pp. 933–945.

Sosa, A., Stoel-Gammon, C. (2006). Patterns of intra-word phonological variability during the second year of life. *Journal of Child Language 33*, 31–50.

Stark, R. (1980). Stages of speech development in the first year of life. In: G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds), *Child phonology 1: Production*. Academic Press, pp. 73–92.

Suzuki, T. N., Wheatcroft, D. & Griesser, M. (2016). Experimental evidence for compositional syntax in bird calls. *Nature Communications*, *7*, 1–7.

Tchernichovski, O. Lints, T. J., Deregnaucourt, S., Cimenser, A. & Mitra, P. P. (2004). Studying the song development process: rationale and methods. *Annals Of The New York Academy Of Sciences*, *1016*, 348–363.

Tchernichovski, O. Mitra, P. P., Lints, T. & Nottebohm, F. (2001). Dynamics of the vocal imitation process: How a zebra finch learns its song. *Science, 291* (5513), 2564-2569.

Vallentin, D. Kosche, G., Lipkind, D. and Long, M. A. (2016). Neural circuits: Inhibition protects acquired song segments during vocal learning in zebra finches. *Science, 351* (6270), 267–271.

Veit, L., Aronov, D. & Fee, M. S. (2011). Learning to breathe and sing: development of respiratory-vocal coordination in young songbirds. *Journal of Neurophysiology*, *106* (4), 1747–1765.

Vihman, M., & Velleman, S. (1989). Phonological reorganization: a case study. *Language & Speech 32*, 149–170.

Warren, T. L. Charlesworth, J. D., Tumer, E. C. & Brainard, M. S. (2012). Variable sequencing is actively maintained in a well learned motor skill. *Journal of Neuroscience*, 32 (44), 15414–25.

Waterson, N. (1971). Child Phonology: A Prosodic View. *Journal of Linguistics, 7*, 179–211.

Wohlgemuth, M. J., Sober, S. J. & Brainard, M. S. (2010). Linked Control of Syllable Sequence and Phonology in Birdsong. *Journal of Neuroscience, 30* (39), 12936–12949.

Yip, M. (2013). Structure in Human Phonology and in Birdsong: A Phonologist's Perspective. In: Bolhuis, J. & Everaert, M., *Birdsong, Speech, and Language: Exploring the Evolution of Mind and Brain*, Cmabridge, MA: The MIT Press, pp. 181-208.

Tables:

Table 1. Terminology for structural units in birdsong and speech

BIRDSONG	SPEECH
Note: a short period of stable (unchanging) acoustic state. Notes are the smallest acoustically distinct units in birdsong.	**Phoneme**: the smallest unit that can contrast word meanings in the sound system of a language. Phonemes are abstract units, and are represented between slashes: /p/ /a/ The realizations of phonemes in speech are termed **Sounds**, and are represented between square brackets: [p] [a]
Song Syllable: continuous sound performed on expiration, followed by a brief inspiratory silent period.	**Syllable**: The minimal unit of organization of sounds. The universal core syllable consists of a vocalic Nucleus, i.e. a vowel, preceded by a consonantal Onset, CV.
Motif/phrase: a short stereotyped sequence of song syllables;	**Word**: the smallest element that can be uttered in isolation with objective or practical **meaning**. A word is thought to be stored with its meaning, grammatical class (noun, verb, adjective, etc.) and sound structure in the Mental Lexicon.

Table 2. Invariant consonant sequences in early Dutch child language (Levelt, 1994)

Target Dutch Word (+translation)	**Phonological representation**	**Child Production**	**Target Transition**	**Produced Transition**
poes (cat)	/pus/	[pus]	Labial /p/- Coronal /s/	Labial [p] - Coronal [s]
soep (soup)	/sup/	[fup]	Coronal /s/ - Labial /p/	Labial [f] -Labial [p]
slapen (sleep)	/slapə/	[fapə]	Coronal /s/ - Labial /p/	Labial [f] -Labial [p]
tekenen (draw)	/tekənə/	[tekə]	Coronal /t/ - Dorsal /k/	Coronal [t] -Dorsal [k]
katten (cats)	/kɑtə/	[tɑkə]	Dorsal /k/ - Coronal /t/	Coronal [t] -Dorsal [k]

Figure Legends:

Fig. 1. Development of birdsong syllables : a, Left, a sound spectrogram (time-frequency plot) of a song of an adult male zebra finch (90 days old). Black lines indicate syllables – bursts of sound separated by brief silent gaps; the song consists of discrete syllable types (indicated by letters), which are repeated in short stereotyped sequences - motifs. Right, distribution of two features characterizing syllable structure (duration and mean Frequency Modulation) for syllables performed by the same bird during an entire day. Discrete syllable types appear as distinct clusters in the distribution. **b**, Spectrogram (left) and syllable feature distribution (right) showing juvenile subsong performed by the bird in **a** at 40 days of age (notations as in **a**). Syllable structure and durations are highly variable, with a broad (un – clustered) distribution. No distinct syllable types (and consequently, no distinct syllable sequences) are observed. **c**, Spectrograms showing the developmental trajectory of two renditions of a proto-syllable of a juvenile zebra finch (bottom) that differentiated into two acoustically discrete syllable types of its target song (top plot). Days from first exposure to tutor song are indicated on spectrograms. Adapted from Tchernichovski et al., 2001; **d**, Distributions of two syllable features (duration and mean goodness of pitch) in a bird trained to perform one syllable (red cluster) early in development; and then exposed to an additional syllable (blue cluster). Day 0, day of first exposure to the new syllable. Acoustic variability is locally regulated within the song, as is evident from the considerable difference in size and rate of shrinking between the two clusters. Adapted from Ravbar et al 2012.

Fig. 2. **The classification of serial order in articulations. a**, The place of articulation for consonants and vowels, and the articulatory organs involved in consonant production., Vowels are categorized into three types according to the horizontal position of the tongue:

front, center and back. Three places of articulations are shown: labial, coronal and dorsal. Labial consonants are mainly articulated by the lips and jaw. Coronal consonants are mainly articulated by the tongue apex and jaw. Dorsal consonants are mainly articulated by the tongue dorsum and jaw. **b**, Serial order in articulation of consonants in consonant-vowel-consonant(-vowel) sequences. (i) Sequences consisting of consonants produced at the same place of articulation. (ii) Sequences produced by movements from more anterior place to more posterior one. Adapted from Oohashi, Watanabe and Taga, 2013.