Framing a socio-indexical basis for the emergence and cultural transmission of phonological systems

Andrew R. Plummer[a,*], Mary E. Beckman[b,*]

[a] *Department of Computer Science and Engineering, The Ohio State University*
[b] *Department of Linguistics, The Ohio State University*

Abstract

Moulin-Frier et al. (2016) proffer a conceptual framework and computational modeling architecture for the investigation of the emergence of phonological universals for spoken languages. They validate the framework and architecture by testing to see whether universals such as the prevalence of triangular vowel systems that show adequate dispersion in the F1-F2-F3 space can fall out of simulations of referential communication between social agents, without building principles such as dispersion directly into the model. In this paper, we examine the assumptions underlying the framework, beginning with the assumption that it is such substantive universals that are in need of explanation rather than the rich diversity of phonological systems observed across human cultures and the compositional ("prosodic") structure that characterizes signed as well as spoken languages. Also, when emergence is construed at the time-scales of the biological evolution of the species and of the cultural evolution of distinct speech communities, it is the affiliative or affective rather than the referential function that has the greater significance for our understanding of how phonological systems can emerge de novo in ontogeny.

Keywords: speech communication, phonological diversity, compositionality, phonological acquisition, language evolution

[*] Corresponding author

Email addresses: plummer.321@osu.edu (Andrew R. Plummer), beckman.2@osu.edu (Mary E. Beckman)

1. Introduction

In their target article, Moulin-Frier et al. (2016) (henceforth, MDSB when referring to the authors) address the issue of the origin of substantive phonological universals – i.e., of cross-linguistic generalizations about the possible (or most frequent) vowel inventories, consonant features, syllable shapes, and so on. Although the premises of the endeavor have sometimes been questioned (see, e.g., Ladefoged, 1983), formulating principles to describe and explain such generalizations has been a key part of the development of phonetic theory over the past half century and more (see, e.g., Jakobson et al., 1951/1969; Greenberg, 1965; Liljencrants and Lindblom, 1972; Macken and Ferguson, 1981; Bell, 1978; Ohala, 1983; Maddieson, 1984; Lindblom et al., 1984; Stevens, 1989, among many others). In this tradition, MDSB build on a long line of investigation that attempts to model the emergence of phonological universals in the species from more general cognitive constraints related to the assumed functions of speech communication in interaction with channel constraints from the sensory-motor system. Specifically, they put forward a conceptual framework and computational modeling architecture, collectively called "COSMO," for simulating the interaction between these two types of constraint in a population of agents that engage in a language game. We focus our commentary by examining the foundational assumptions that are the basis for the conceptual framework.

Within the COSMO framework, speech communication is, at its most essential level, taken to be "the modification of the internal-knowledge state of a listener, by a speaker, through the use of communication stimuli" (Moulin-Frier et al., 2016, p. 2 (ms)). Accordingly, the agents carrying out speech communication are taken to possess speech perception and production capabilities, along with the cognitive capacity to integrate the sensory and motor knowledge under-girding their roles as both listener and speaker. Several of the substantive universals that characterize spoken language phonologies are hypothesized to emerge from speech communication among agents, under the condition that their communication obeys a set of sensorimotor and cognitive constraints. MDSB take the three main constraints to be the following:

1. *adequacy* – "[the agents] must select adequate communication stimuli that are reasonably easy to produce and process,"
2. *parity* – "a good correspondence must be ensured between the speaker's motor repertoire and the listener's perceptual repertoire,"

3. *reference* – "[the agents] must know the correspondence between these motor and perceptual repertoires and the objects in the external world." (Moulin-Frier et al., 2016, p. 2 (ms))

While MDSB state that, "[i]n its general form, [the conceptual framework] is agnostic in relation to choices about adequate phylogenetic precursors of these requirements" (Moulin-Frier et al., 2016, p. 6 (ms)), they do make specific choices about how to implement them within the chosen computational architecture as they set about demonstrating that the framework is capable of characterizing, reasoning about, and explaining a variegated set of phenomena, including consonant-vowel co-occurrence trends that might originate in the same pre-verbal mandibular oscillation that supports the emergence of the syntagmatic differentiation between consonants and vowels within syllables; the rarity of pharyngeal consonants relative to labials, coronals, and dorsals across languages; and the interaction between motor and auditory feedback in the emergence of an adequately dispersed vowel system.

At first glance, it may seem that situating this set of phenomena within a single conceptual framework that can be "agnostic" about phylogenetic precursors constitutes an important breakthrough in modeling. After all, this approach of making broad generalizations to manage the notorious degrees of freedom problem enabled such monumental achievements as Quantal Theory (Stevens, 1972), the notion of Adaptive Dispersion (Liljencrants and Lindblom, 1972), and the Frame-Content Model (MacNeilage and Davis, 2000).

Yet, a closer look in comparison to other, less ambitious models leads us to ask whether it is time now to instead delve more into the details. For example, by contrast to the models in Ishihara et al. (2009), Miura et al. (2012), and Rasilo et al. (2013), among others, there is no sense in which the COSMO model applies to mother-infant dyads (or other asymmetrical teacher/learner groups) interacting on the short-term scale of ontogeny. Nor is it simulating interactions in a population that gradually changes as older teacher agents die off and new learner agents are born and then grow into teacher agents themselves, as in de Boer (2000). As noted by Kirby and Hurford (2002), Vogt (2005), Beckman and Edwards (2010), Chater and Christiansen (2010), and others, a full explanation of the emergence of language universals will require integration of their characterizations at the levels of ontogeny and cultural mutation and phylogeny, as well as models that better capture how the emergence of characteristics at one level can affect and be affected by the emergence of characteristics at another level. These three characterizations

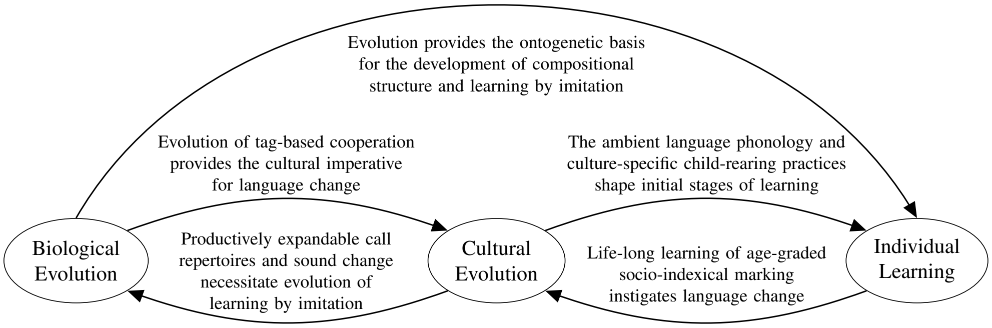
Evolution provides the ontogenetic basis
for the development of compositional
structure and learning by imitation
Evolution of tag-based cooperation
provides the cultural imperative
for language change
The ambient language phonology and
culture-specific child-rearing practices
shape initial stages of learning
Biological
Evolution
Productively expandable call
repertoires and sound change
necessitate evolution of
learning by imitation
Cultural
Evolution
Life-long learning of age-graded
socio-indexical marking
instigates language change
Individual
Learning

Figure 1: The chain of influences relating biological evolution, cultural evolution, and individual learning factors needed in characterizing the evolution of the human capacity for language and the pathways for language learning and change. See Sections 2 and 3 for explication of factors labeling each linking arc.

require reference to vastly different time-scales, differing social functions and agent states within and across these time-scales, differing environmental contexts, etc., as suggested in Figure 1.

In section 2 we elaborate on this point by evaluating MDSB's formulation of agents and their communicative interactions against what is known about the evolution of language at the three different time-scales in Figure 1. We focus especially on an evaluation at the time-scale of ontogeny, which is the best understood of the three levels. We also evaluate the extent to which the most studied phonological universals are challenged by the existence of languages such as American Sign Language, while noting a phonological generalization that seems to hold even in the very recently evolved Al-Sayyid Bedouin Sign Language (see, e.g., Sandler et al., 2005). Explicitly, this generalization is that all human languages provide prosodic mechanisms for sequencing and grouping sequences of phonological patterns, such that an infinite number of potentially very complex larger patterns can be produced and interpreted in terms of reusable smaller parts.

In section 3, we then review research that reveals potential starting points for investigating how such prosodic mechanisms develop in human infants who have normal hearing and are born into speech communities where they have regular opportunities to engage socially with older community members who speak around and at them. The time course for these developments, which are the ontogenetic bases for duality of patterning, place them well before typically-developing infants begin to engage in the triadic "deictic game" of joint attention to a shared referential object that the COSMO framework takes to be the primary form of interaction between social agents. Rather, these developments seem to be initially supported by infants engaging with caregivers in a simpler dyadic "imitation game" where vocalizations within the context of the game encode affiliative information that infants use to relate representations of the participating agents.

Altogether, the research reviewed in this commentary raises doubts about the manner in which MDSB construe adequacy, parity, and reference as a joint initial basis for the emergence of phonological systems de novo, while simultaneously suggesting alternative conceptualizations of emergence at each of the three time-scales. While a full characterization of the emergence of phonological systems at any of the three time-scales is beyond the scope of any current modeling framework, enough is now known about how contingent dyadic social interactions can shape infants' vowel-like vocalizations in the first 5 months of life that a framework for modeling the emergence of phonological systems in ontogeny might well begin by modeling this process. We conclude with a summary of our arguments, stressing the need for renewed attention to results from the experimental literature on early infancy in developing social agent models of the emergence of phonological systems at the time-scale of ontogeny without imputing the capacity for reference at an inappropriate place in the time course of normal human development.

2. Reconsidering the universality of language forms

In the introduction section of their target paper, MDSB motivate the simulations that are described in sections 6.2-6.4 as a test of a particular instance of a class of explanations for the observation that "human languages display a number of regularities" which make the rich diversity of human languages "appear merely as variants of a single system" – i.e., as a test of whether the COSMO framework can provide a plausible model of the evolutionary circumstances that led to the "universality of language **forms**" (Moulin-Frier et al., 2016, p. 1 (ms), emphasis added). More specifically, the goal is to design a community of social agents and a type of communicative interaction between agent pairs in a model that explicitly instantiates their assumptions about parity, adequacy, and reference, so that simulations can be run with random pairs of agents assigned to play the speaker role or the listener role in a series of social interactions that continue until a shared repertoire of referential “language forms” emerges. These model outcomes are then examined to see whether they display a particular regularity, such as a tendency for the set of vowel sounds [i], [e], [a], [o], and [u] to emerge as the repertoires of five referential forms that are simulated in section 6.2. If an expected tendency is observed, then this result is interpreted as evidence that the design characteristics of the model can explain the regularity without building the regularity into the system as an explicit design principle. The advantage of this kind of simulation is that the evaluation metric can be a quantitative one, such as a comparison between the proportion of simulations that show the universal pattern and the proportion of languages in the UPSID database of phoneme inventories (Maddieson, 1984, 1991) that show it. In this section, we discuss two other, more qualitative evaluation metrics.

The first involves the reference frames and granularity of representations of the phonological universal. The explanatory power of the simulations is diminished if the reference frames for generating and perceiving the forms and the granularity of the representations of the forms in the shared repertoires that emerge in the model communities do not match the reference frames and the granularity of the representations used in the language descriptions that give rise to the observed universal. The three sets of simulations in the target paper show varying degrees of mismatch, ranging from a major discrepancy (for the simulations in sections 6.2 and 6.4) to a seemingly insurmountable incomparability (for the simulations in section 6.3). In all three sets of simulations, the reference frames are combinations of VLAM (Boë and Maeda, 1998) articulatory parameter settings and associated cross-sectional area functions simulating the adult female vocal tract of a social agent producing “words” (i.e., the repertoire of referential “language forms” of the community of social agents), and the grain of the representations of these forms is the resolution in Barks of the resulting points or trajectories within the F1-F2-F3 maximal formant space (MFS) for the vocal tract used in the model. By contrast, the three phonological universals to which the model outcomes are compared are observed trends in transcribed consonant and vowel inventories in the UPSID database (for the simulations in sections 6.3 and 6.2) or in transcribed consonant-vowel sequences in babbling vocalizations, first words, and dictionaries (for the simulations in section 6.4).

To establish a correspondence between such disparate reference frames for the consonants in isolation that are the “words” in section 6.3 and the consonant onset points of the CV syllables that are the “words” in section 6.4, MDSB overlay 8 ellipses on the F2-F3 plane of the adult female MFS and equate the centers of these ellipses with the IPA symbols for labial, dental, alveolar, palatal, velar, uvular, pharyngeal, and epiglottal stops. And to establish a correspondence for the vowels that are the "words" in section 6.2 and the vowel ending points of the CV syllables that are the "words" in section 6.4, they subdivide the F1-F2 plane of the adult female MFS into 7 regions that are then equated with the vowel symbols [i], [e], [ɨ], [ə], [o], [u], and [a]. In other words, for the simulations in section 6.3, the correspondence is one that maps to a set of symbols for transcribing a type of sound that rarely or never occurs as the sole segment in a word in any of the languages in the UPSID database (i.e., stop consonants) from a set of specifications for static points with very low F1 in the MFS of the model speaker (i.e., sounds which observer phoneticians might transcribe as very peripheral vowels).

Moreover, even for the vowel sounds and consonant-vowel sequences that can occur as isolated word forms in many spoken languages (and that phonetician observers might transcribe as a representation of an infant's prelinguistic vocalizations), the correspondence is one that maps directly to the symbol set from the MFS for one type of VLAM vocal tract without first establishing the correspondence between that MFS and the MFS for any other type of vocal tract, including vocal tracts that are more appropriate for modeling the productions of the infants and toddlers whose babbling vocalizations and first words were transcribed to make the databases that MacNeilage and Davis (2000) offer as a part of the evidencefor the universal in section 6.4.

Thus, the COSMO framework simulations depend on there being a set of symbolizable phonological units O_L that can be inferred directly from the speech signal by the "auditory processing capabilities" of a purportedly prelinguistic (and hence pre-phonemic) social agent whatever the circumstances – i.e., whether the agent is listening to auditory stimuli generated by another agent's vocal tract (as when playing the role of listener in dyadic or triadic interactions such as the "imitation game" or the "deictic game") or to auditory stimuli generated by the agent's own vocal tract (as in solitary babbling or when playing the role of speaker in dyadic or triadic interactions). That is, the metric depends on there being a direct mapping between representations of specific vocalizations in a talker-specific acoustic reference frame and representations of language forms in a tool that was developed for doing linguistic fieldwork. The gross discrepancy between these levels of granularity can be appreciated by the fact that even the author of UPSID cautions that "in using the database,...[t]he questions examined must be tailored to match the level of detail that is available" (Maddieson, 1991, p. 197). For example, because of the practice of listing only features of "the most basic allophone" of any phonemic category and the "difficulties in deciding cross-language equivalences between places of articulation" the UPSID should not be used to evaluate the relative likelihood that a stop phoneme will be alveolar versus dental (Maddieson, 1991, p. 198). Phonologists who do not axiomatically assume categories such as [d̪] versus [d] as universal innate structures emphasize even more the incommensurability of these two reference frames (see, e.g., Ladd, 2014).

The second qualitative evaluation metric involves the assumptions that guide the design of the agents and their interactions and their relationship to the posited time-scale for the emergence of each of the three universals addressed by the simulations in sections 6.2 through 6.4. The explanatory value of the simulations is vitiated if the design characteristics of the social agents and their interactions are counterfactual or implausible for the intended time-scale.

Although MDSB do not explicitly specify the intended time-scale, since the social agents in the COSMO framework simulations are described as being "initially constrained by a set of **prelinguistic** abilities" (Moulin-Frier et al., 2016, p. 2 (ms), emphasis added), and since the repertoires that evolve are extremely small (3 or 5 language forms in total), it might seem at first glance that the models could be simulating the emergence of an initial repertoire of referential vocalizations in the infant with normal hearing in interaction with other members of the infant's initially small social circle – i.e., emergence at the time-scale denoted by the rightmost node in Figure 1. A closer examination of the design characteristics makes clear that this cannot be the intended time-scale for any of the simulations, for two reasons.

First, the social agents in the COSMO-framework simulations are equal and interchangeable. That is, the agents are all endowed with the same vocal tract (which is the size and shape appropriate for an adult female human) so that the speaker agent's motor repertoire can be mapped directly onto the listener agent's perceptual repertoire in the "internalization" of the communication process because of the trivial equivalence between the auditory stimuli S that are generated by the motor gestures M specified for the speaker agent's vocal tract and the auditory stimuli S that would be generated by the internal model of the articulatory-to-acoustic transformation specified for the listener agent's vocal tract. Moreover, every agent can be placed either in the listener role or in the speaker role, and for the latter role, every agent "is provided with a brain that includes a set of cognitive control processes acting on a vocal tract through different articulators" (Moulin-Frier et al., 2016, p. 7 (ms)), so that every social agent not only can map every other social agent's auditory stimuli onto the target agent's motor gestures when playing the listener role, but also already has the motor control to perform the motor gestures that will reproduce these auditory stimuli when it comes time for the agent to play the speaker role. In short, all agent pairs are pre-endowed with "parity" and "adequacy" and the speaker and listener roles are trivially interchangeable.

By contrast, in the social interactions that are critical for the acquisition of a first spoken language, one of the agents already has linguistic abilities, and the other agent — namely, the infant agent, who is the one who has only **pre**linguistic abilities — has a vocal tract with a very different size and shape (see, e.g., Crelin, 1969, 1987; Vorperian et al., 2009; Barbier et al., 2012, 2015). Moreover, the prelinguistic abilities of the infant agent do not yet include anything like the "cognitive control processes" of the other agent. Rather, the ability to produce vowel-like utterances that use the full MFS for a given age is something that only develops over the first year or so of life (see, e.g., Kent and Murray, 1982; Ishizuka et al., 2007; Rvachew et al., 2008)

Second, in the COSMO framework simulations, each member of an interacting pair already "must know the correspondence between these motor and perceptual repertoires and the objects of the external world" so that the communication system that emerges from a sequence of interactions is a repertoire of shared "linguistic forms" that were assigned a referential function by all of the social agents already at the "prelinguistic" beginnings of the sequence. However, the joint attention that supports functional reference in the human infant does not begin to emerge until months after the age when language-specific perceptual processing of vowels and consonants first begins to emerge, and "knowledge of the correspondence" between the infant's vocal motor schemes and "the objects of the external world" is still very rudimentary at an age when the vocalizations produced by infants begin to reflect the influence of the specific languages to which they have been exposed (see literature reviewed in Vihman, 2014, chapters 2–4).

Thus, to be a model of the emergence of a phonological system in the human infant at the shortest time-scale depicted in Figure 1, the social agents would need to be re-designed so that pairs of agents are not initially equal and interchangeable. One member of each pair would be endowed only with the prelinguistic abilities that allow the human infant to develop the correspondences between the infant's own prelinguistic motor repertoire and prelinguistic perceptual repertoire that are exercised already at what Oller (1980) calls the "phonation" and "goo" stages and then expanded at the "canonical babbling" stage. The other member of each pair would need to be endowed with the linguistic abilities to map her representations of the infant's vocalizations onto her language-specific motor repertoire and perceptual repertoire, and to respond in some (potentially quite culture-specific) way that fosters the prelinguistic agent's drive to map from the infant's representations of the other agent's productions onto the infant's vocal repertoire. Also, the "deictic game" of the COSMO framework simulations would need to be redesigned to simulate the patterns of pre-referential social interaction between an agent that is endowed with linguistic abilities and an agent that is endowed with only prelinguistic abilities. It does not matter exactly which patterns these are, but ideally they should be ones that have been documented in some culture where enough careful observation has been done to be able to simulate the interactions without counterfactually endowing the infant agent with cognitive abilities which have not yet begun to emerge at the stages where language-specific perceptual-motor mapping begins to emerge in the infant with normal hearing. We will return to these points in Section 3 after discussing two other places in the schematic in Figure 1 where it might be more appropriate to use the deictic game to simulate interactions between interchangeable agent pairs with only prelinguistic abilities.

One place where this use of the deictic game might be appropriate is in modeling the evolution of language in a newly created community of potential language users who do not yet have language because they were deprived of speech input in infancy. While such communities are not the normal environment for the emergence of language at the time-scale of cultural evolution (i.e., the middle of the three nodes in Figure 1), they are attested and at least one has been studied since inception. This is the Deaf community in which Nicaraguan Sign Language has emerged. As Meir et al. (2010, p. 267) point out, Nicaraguan Sign Language and other emerging sign languages provide "a natural laboratory for studying the development of linguistic structure and its interaction with the nature of the language community." The phonological systems of such new languages share several important phonological properties with more established sign languages such as ASL and also with spoken languages. These properties are the conventions that come to be established for sequential and simultaneous "prosodic" structure. That is, even the newest sign languages have conventions for lawful sequencing of two or more elemental language forms in a primary (manual) gestural stream (see, e.g., Senghas et al., 2004) and conventions for grouping the manual gestures into clearly demarcated phrases both via the manual analogs of final lengthening and silent pause and via the superposition of simultaneous "suprasegmental" gestures of the signer's eyes, mouth, and torso (see, e.g., Sandler et al., 2005). These conventions differ from language to language even after taking modality differences into account, and they are clearly part of the phonological competence that can differentiate native speakers from non-native speakers of a sign language as well as of a spoken language (see, e.g., Braem, 1999).

Of course, newly emergent sign languages will not have vowels, consonants, and CV syllables, and the agents' perceptual and motor repertoires would need to be recast for the manual-visual channel in order to be able to simulate the emergence of language in interactions between interchangeable agent pairs with only prelinguistic abilities. However, the presence of rhythmic and "suprasegmental" prosody in even these youngest human languages suggests a deep phylogenetic origin for the capacity for sequencing and prosodic structuring of sequences. So these capacities seem important characteristics to try to model in simulations of the emergence of phonological universals, and emerging sign languages give us a way to model their evolution at the middle of the three time-scales in Figure 1.

More generally, perceptual sensitivity to prosodic structure for spoken languages emerges very early in infancy (see, e.g., Mehler et al., 1988, Nazzi et al., 1998) and the ethnographic literature often describes differences in response to infants' "babbling" in a way that suggests that this type of non-reflexive, more speech-like vocalization begins to be produced in the first year of life at a schedule that is somewhat impervious to cultural variation in the proclivity of adults to assign it referential function (see, e.g., Blount, 1972; Ochs and Schieffelin, 1982; Richman et al., 1992). So simulating the emergence of prosodic structure also in the normal ontogenetic progression for hearing infants could give us insights into the interaction of phylogeny and ontogeny in the longest arc at the top of Figure 1. We return to this point in section 3.

The other place where it might be appropriate to use the deictic game to simulate interactions between interchangeable agent pairs with only prelinguistic abilities is in models of the evolution of a vocal communication system with functional reference at the longest time-scale in Figure 1. As MDSB note, there is evidence of functional reference in the vocal communication systems of many non-human animals. And the spectrograms illustrating typical tokens of the different call types in the repertoires of these species suggest that "Adaptive Dispersion" – i.e., the principle that vocal communication systems should maximize acoustic distance to enhance distinguishability of forms – is a very general naturally emergent outcome of the evolution of sound systems at this time-scale. So it seems likely that, with the substitution of species-specific anatomical models and species-appropriate adjustments to the auditory model, the COSMO framework could be used to simulate the evolution of the acoustically differentiated alarm calls of chickens (see, e.g., Evans et al., 1993), meerkats (see, e.g., Manser et al., 2001; Hollén and Manser, 2007), vervets (e.g., Seyfarth et al., 1980), and so on. Of course, in these other species, the principle of maximizing auditory distance characterizes the emergence of a small biologically transmitted repertoire of vocalizations which co-evolved with the social organization of the species as a whole. As Seyfarth and Cheney (2003) Tanaka et al. (2006, p. 8), Griebel and Oller (2008, p. 11), and others point out, such systems tend to be "conservative", having stereotypic pan-species forms with a fairly fixed form-to-function mapping that is modified minimally in development.

It is important to note that such systems are very different from the diverse, enormous, and productively expandable repertoires of vocal signs that constitute the vocabularies of human spoken languages. Indeed, by comparison to the vocal communication systems even of the other apes, the human vocal communication system shows an enormous diversity in the repertoire of vocal signs across different groups. For example, in his analysis of vocalization rates in the Kibale community of chimpanzees, Arcadi (2000) was able to apply the same inventory of vocalization types that was catalogued by Goodall (1986) for the Gombe community. And while there is emerging evidence of a capacity for vocal learning in studies of inter-community differences in finer-grained acoustic characteristics of some of these vocalization types, especially the long-range vocalizations (pant hoots) that are the most commonly recorded call type for adult males (see, e.g., Arcadi, 1996; Crockford et al., 2004), pant hoots are observed in all chimpanzee communities that have been studied to date, and they contrast both in general acoustic shape and social function to other common vocalizations such as the pant grunts that are addressed to dominant conspecifics. By contrast, different human languages typically have different wordforms for different referential functions (i.e., different units at the lexical level that Martinet, 1949, called the “primary articulation” of the speech stream). Moreover, they show such lexical differences not only for the kinds of referent that tend to differ across cultures (tools and other material artifacts, food-preparation practices, ritual behaviors, etc.), but also for pan-species referents such as body parts, kin relationships, the physically necessary behaviors of eating, sleeping, etc. Indeed, such lexical differences often are salient characteristics differentiating dialects within a single language. This diversity of vocalization repertoires is a critically important aspect of the cultural diversity that characterizes the species, making vocal learning a necessary capacity for any human infant that is born into a culture that is associated with a spoken language.

Moreover, different dialects of a spoken language, as well as different languages, typically have different inventories of phonemes (i.e., different units at the sublexical level that Martinet, 1949, termed the “secondary articulation” of the speech stream) and even when phoneme inventories are similar, pronunciation details can differ radically, as shown, for example, by Nartey (1979), Ladefoged and Bhaskararao (1983), and Disner (1983), among many others. When human infants are born into a culture associated with a spoken language, they learn not just the word forms specific to the language or dialect, but also the language- or dialect-specific phonetic details of the component vowels and consonants. So, while toddlers growing up in Greek- and Japanese-speaking homes both learn the “same” modal 5-vowel system, the toddlers growing up in Greek-speaking homes learn the less back /a/ and more rounded /u/ of Greek rather than the /a/ and /u/ of Japanese (Chung et al., 2012). Similarly, while toddlers growing up in Swedish- and English-speaking homes both learn the “same” modal 3-place voiceless stop system, the toddlers growing up in Swedish-speaking homes learn the more dental, less aspirated /t/ of Swedish rather than the /t/ of English (Stoel-Gammon et al., 1994).

In short, when compared to vocal communication systems with referential function in other social animals, such as the repertoire of alarm calls in vervets, what stands out for human spoken languages is not the broad-stroke substantive generalizations that might be made when comparing transcribed vowel and consonant inventories in UPSID. Rather it is the extremely rich diversity of referential repertoires across speech communities and the incommensurability in details of pronunciation even for forms that might be transcribed as the identical sound shape. To understand the emergence of a fully human repertoire of referential vocalizations at this longest time-scale, therefore, it seems essential to try to identify the evolutionary forces that created both this rich diversity of repertoires across communities and the associated imperative for robust ontogenetic processes for vocal learning (see Baronchelli et al., 2012 for one of the many other papers that have made the same point, but with respect to morpho-syntax rather than to phonology).

The relevant phylogenetic precursors for these two universal properties of human communication systems are less likely to be found in conservative systems such as alarm calls than in systems that show evidence of vocal learning such as patterns of convergence between individuals who cooperate to defend shared territory or to feed each other's children (and complementary patterns of divergence between groups that compete for resources). Evidence of vocal learning in chimpanzee pant hoots can be seen in patterns of convergence between socially bonded pairs of males within a community (Mitani and Gros-Louis, 1998) (as well as in the patterns of divergence across communities already noted above). The fact that the convergence is seen in adult males rather than females is related to the fact that non-kin social bonding and cooperation is more characteristic of adult males in this species (see literature reviewed in Mitani, 2009). Other primate systems that show evidence of vocal learning in both sexes as well as in juveniles include the affiliative calls of gibbons (Mitani, 1985; Haimoff, 1986) and pygmy marmosets (Elowson et al., 1998; Snowdon and Elowson, 1999). These are both species that show complex patterns of group formation, with associations between mated pairs and also between adults and unrelated subadults for cooperative actions including allo-care of infants after weaning (Brockelman et al., 1998; Snowdon and Cronin, 2007; Elowson et al., 1998). In this context, it seems noteworthy that Clarke et al. (2006) describe gibbon affiliative calls as showing an internal compositional "syntax" that is very reminiscent of Martinet's principle of "double articulation" (i.e., the principle that contrasting units at the "primary articulation" level are composed of unique sequential combinations of units at a "secondary articulation" level), and that it is this "syntax" that facilitates the emergence of pair-specific and larger family-group specific variants of the calls. If call structure design interacts with vocal learning in this way, perhaps community-wide changes to the repertoire of calls with a more clearly referential function (such as alarm calls or food calls) also could emerge, to provide further ways of rewarding cooperation by the selective transmission of knowledge only to in-group individuals.

The link between vocal learning and patterns of cooperation is reminiscent of the function that Cohen (2012) proposes for "accent" – i.e., the community-specific pronunciation patterns that are flexibly acquired by human children but difficult to remold in later life. Cohen suggests that accent functions as a reliable mechanism for tag-based cooperation among flexibly large and diffuse groups. The ability of human infants to acquire the accents of groups that vary substantially in size and diffuseness facilitates potential cooperation across a broad range of environmental scenarios. In human infants, the onset of canonical babbling also provides the ontogenetic basis for the emergence of consonant-vowel sequences and the "double articulation" of spoken language phonologies.

Could COSMO-framework simulations be used to develop and test models of the relationship between the emergence of accent in the evolution of human social groups and the emergence of canonical babbling and the subsequent decomposition of the canonical syllable into recombinable consonants and vowels in the evolution of human vocal learning? Could such simulations also shed light on the evolution of ontogenetic changes in the human vocal tract, which Crelin (1987) and de Boer (2010) suggest are adaptive for the production of robustly differentiated vowels, albeit maladaptive for safeguards against Sudden Infant Death Syndrome and asphyxiation during feeding? In the next section of this paper, we try to establish a basis for addressing such questions by reviewing what is known about speech-like vocalizations in human infants and discussing some of the suggestions that have been made about phylogenetic precursors for the evolution of the physical and motor developments that support the emergence of simultaneous and sequential compositionality at this shortest time-scale in Figure 1.

3. The ontogenetic bases for compositionality

As noted in the previous section, the literature on emerging sign languages suggests that the capacity to use prosody to impose compositional structure on one's own and others' communicative gestures is the one irrefutably universal phonological generalization. What are the design requisites of potentially illuminating models of the emergence of this phonological universal in infants who are acquiring spoken languages? In this section, we address this question by pointing to two landmarks in the emergence of the potential for word-level compositionality in the infant's vocal development over the first year of life, noting the literature on relevant concurrent anatomical and physiological changes and on the social function or context of use of the infant's vocalizations as well as any potential phylogenetic precursors that have been identified.

The second and better studied of the landmarks is the emergence between 6 and 8 months of what Oller and Eilers (1988) term “canonical babbling” or “canonical syllable” – a type of vocalization that involves articulatory movement superimposed on the controlled exhalation phase of the respiratory cycle that momentarily interrupts the airflow in such a way as to give rise to the percept of an alternating series of a stop- or nasal-like consonant followed by a vowel. Kent (1984) and many others have noted that this landmark is tied to the development of rhythmicity in general; it occurs in association with peaks in stereotypic rhythmic movements of the fingers, limbs, and torso, and the timing of its onset seems to be fairly impervious to differences in socioeconomic circumstance and mild degrees of hearing impairment (see, e.g., Nathani et al., 2007, and literature reviewed there). Relatedly, MacNeilage et al. (1997, p. 269) have characterized this type of vocalization as “a syllabic frame produced by an open-close mandibular oscillation” which is coordinated with and subdivides a phonatory gesture. That is, the defining characteristic of canonical babble is not an active sequential coordination of a controlled consonant constriction gesture with an independently controlled vowel posture, but instead a simpler rhythmic movement of the jaw which initially is combined with static postures of the tongue and lips, so that the decomposition of the vocalization into a consonant segment followed by a vowel segment is an artifact of the acoustically abrupt transition between the aerodynamic regimes for a closed and then an open oral cavity.

The significance of this landmark in language socialization routines seems to vary across cultures. In some cultures, including the ones in which early mother-infant interactions are best studied, the emergence of the syllable-like rhythm promotes “child-centered accommodation” (Ochs and Schieffelin, 1982). This is a style of social interaction in which the infant’s caretakers impose a word-like compositional structure on the infant’s vocalization by imputing a referential intent to the infant and then responding in some way, such as repeating the inferred word in an elaboration or question, that facilitates the infant’s bootstrapping from babbling into first word productions (see, e.g., Snow, 1977, and other work that builds on this early study of “motherese”). Thus, in some cultures, the caretaker’s “rich interpretation” of the infant’s babbling seems to accelerate the transition from the purely affiliative or affective “imitation game” exchanges of dyadic mutual attention to the “deictic game” exchanges that become possible when the infant is drawn into triadic joint attention to an external referent. Work reviewed by Ochs and Schieffelin (1982) and Kulick and Schiefflin (2004) suggests that in other cultures, an infant's canonical babbling is not assigned such "rich interpretation" and either "child-centered accommodation" happens only after the infant has begun to produce semi-intelligible referential speech without the explicit re-modeling of the imputed intent, or all "accommodation" is a more didactic "situation-centered" modeling of situation-appropriate utterances that the caregivers present to the infant to imitate. Moreover, even in cultures where canonical babbling marks the infant's debut as a conversational partner in deictic game exchanges, there is substantial inter-individual variation in how the infant then develops enough motor control over tongue tip, lower lip, and other articulators to have a less constrained"content" for the CV frames, and there is equally substantial inter-individual variation in how the infant elaborates on the simple CV frame to be able to produce the other types of syllables and larger foot-level structures that are characteristic of the words of the ambient language that the infant is learning how to say. However, despite this early variability, early word productions and concurrent babbling routines of toddlers who are acquiring the same first language do tend to become more and more alike over time, and also more differentiated from those of toddlers acquiring a different first language, so that the child's phonology eventually converges on the shared pronunciation norms of the speech community (see, e.g., Vihman, 1993). Thus, while different cultures may differ in whether "rich interpretation" and "child-centered accommodation" are the normal response to the emergence of canonical babble, the utterance-internal sequential prosody that is enabled by the alternation between aerodynamic regimes for closed and open oral cavity in this type of infant vocalization provides an ontogenetic basis for the phonological generalization that, in all spoken languages, words have a productive internal serial structure – i.e., that a variety of consonant constriction gestures and vowel postures can be rhythmically coordinated with a single exhalation-phase phonation gesture in the creation of novel sequences to productively expand the repertoire of language forms.

To better understand the relationship between the rhythmicity and the eventual productivity that it enables, we compare canonical babbling to three other types of primate signal that have internal sequential organization. The type that shows the clearest analog to the internal serial compositionality of spoken words is the call repertoires of gibbons. There do not seem to be anatomically-based models of the production mechanisms as of yet, but the descriptions and figures illustrating gibbon call acoustics suggest a fairly simple sequencing of exhalation-phase phonation gestures with high tonality.

That is, the internal structure seems to be a series of more or less musical "notes" that are differentiated from each other by their durations and fundamental frequency trajectories and separated by brief pauses for inhalation. The calls of wild white-handed gibbons are composed by concatenating these "notes" in different ways to differentiate context-specific call types (see, e.g., Clarke et al., 2006). In white-handed gibbons, Simangs, and many other gibbon species, different sequences of "notes" also individuate the "duets" produced by mated pairs and other close social partners (see, e.g., Geissmann, 1999, and other studies cited there). And it seems to be the rhythm of the notes and not just the sequence pattern that converges in the duets of black-handed gibbon mother-daughter pairs observed by Koda et al. (2013). However, gibbon calls seem relatively slow by comparison to the articulation rate of human speech, with all but the fastest note sequences (involving the short "wa" notes) in the illustrative figures in the study by Clarke et al. (2006) showing a longer period than the average canonical syllable durations measured in studies such as Levitt and Wang (1991). Thus, contra Hockett and Ascher's (1964) suggestion, gibbon calls do not seem the most likely homologue to the basic physical mechanism for sequential compositionality in human spoken languages.

Looking next to a more closely related species, we find a different rhythmic mechanism in the build-up phase of chimpanzee pant hoots, a call-internal repetitive structure which Fedurek et al. (2013) suggest evolved in order to be sustained over variable durations, so as to facilitate recruitment of other conspecifics into the "chorusing" bouts that function to develop and maintain adult male chimpanzee social bonds. This serial rhythm is created by the rapid alternation of inhalation- and exhalation-phase phonation gestures, possibly in coordination with an alternation between a more open and a less open mouth. The rhythm of phonating on both the inhalation and exhalation phase of the respiratory cycle is characteristic also of the pant call and the pant-grunt call that are used in face-to-face greeting and grooming by juveniles and adults of both sexes (see description and references listed in Table II of Parr et al., 2005) as well as of the laughter vocalization that infant chimpanzees emit in bouts of tickling and other turn-taking play with their mothers (Ross et al., 2009). However, the rhythm of the pant hoot lead-up phase seems somewhat slower than that of canonical babble, and unlike in the gibbon calls, none of the inhalation and exhalation segments except for the last extended "climax" ones (i.e., the "hoot" part of the call) seem to be productively combined with other independently specified properties, such as variable melodic trajectories, to productively expand the call repertoire.

Finally, the gesture that has been discussed most often as a phylogenetic precursor to canonical babble is not a vocal signal but instead the facial signal of teeth-clacking or lip-smacking. This is an oral gesture without accompanying phonation that wild chimpanzees produce in the context of grooming (see Parr et al., 2005, and references there). A similar gesture has been observed also in at least four groups of captive or semi-free-ranging groups of rhesus monkeys, where its biomechanics and temporal structure have been studied extensively (Morrill et al., 2012; Ghazanfar et al., 2012; Ghazanfar and Takahashi, 2014), along with observations of its context of use. Notable results from this literature are that adult rhesus monkeys produce the gesture in face-to-face affiliative encounters that do not involve grooming as well as in the context of grooming (Hinde and Rowell, 1962), that mother-infant pairs produce lip-smacks in affiliative exchanges during sustained mutual gaze (Ferrari et al., 2009), that neonate rhesus monkeys readily imitate the lip-smack gesture when it is modeled by a human experimenter (Ferrari et al., 2006), and that the speed of the lip-smacking gesture tends to become less variable and to increase with age, reaching a fairly stable 5 Hz oscillatory rhythm at maturity that has been compared to that of the jaw in human speech (Morrill et al., 2012). However, the lip-smacking gesture is not combined with a phonation gesture, much less with a complex melody or with any other potentially independently specifiable acoustic property that can be harnessed to productively expand the signal repertoire.

In summary, while we know of no primate vocal communication system other than human spoken languages that subdivides phonation gestures using a rhythmic alternation between potentially independent oral gestures to alternately obstruct and then allow airflow in coordination with a sustained phonation gesture that is produced on just one phase of the respiratory cycle, there are intriguing parallels in the affiliative signals of at least three other primate species. Thus, the use of call-internal rhythmicity in socially charged affiliative exchanges is clearly not unique to humans. Therefore, regardless of whether the lip-smacking gesture is a homologue for the rhythmic aspect of canonical babble, as MacNeilage and Davis (2000) and Morrill et al. (2012) suggest, or is merely another very suggestive analogue, this literature does suggest that we might find homologues in affiliative exchanges for the other aspect of phonological compositionality that is realized in synchrony with the mandibular rhythm to produce the hallmarks of canonical babble.

This other aspect of phonological compositionality is the most truly universal one. It is the simultaneous compositionality that is manifest even in newly emerging sign languages in the co-production of “suprasegmental” (facial) prosody with “segmental” (manual) gestures, and there are clear homologues in many other primate species. That is, many primate calls combine independently controlled dynamic laryngeal gestures with variable (static) oral gestures to coproduce contrasting pitch and timbre patterns in ways that are independent of the backdrop timbre cues to overall vocal tract size. For example, the coo call that is the characteristic greeting / spacing call of many species of macaque combines a variable pitch pattern with the filter-timbre property of protruded lips, which contrasts with the filter-timbre property of the bared teeth threat call. In coo call exchanges observed in wild populations of Japanese macaques, variation in the duration and overall fundamental frequency height of the initiating coo call is associated with distance from potential recipients, whereas variation in the call-internal fundamental frequency trajectory is associated with the individual monkey initiating the exchange (Sugiura, 2007). Moreover, the call-internal trajectory can be productively varied, as demonstrated in an ingenious study by Sugiura (1998) in which he recorded coo calls from seven monkeys in two different wild populations and then selected calls to use as stimuli in a playback experiment several months later. On each trial, the target monkey was presented with a coo call stimulus from her most frequent coo-call exchange partner and her coo calls after hearing the stimulus were recorded and the first call recorded on each trial was classified as a response call or a random call depending on its latency from the onset of the stimulus call. Comparing the acoustics of the elicited response coos versus random coos to the acoustics of the stimulus coos, Sugiura found that duration and maximum fundamental frequency were positively correlated both in the stimulus- / response-coo pairs and in the stimulus- / random-coo pairs, but the fundamental frequency trajectory measures (the time from call onset to the point of peak fundamental frequency and the frequency excursion size) were positively correlated only in the stimulus- / response-coo pairs.

In human infants, this capacity to combine varying melodies with different timbre properties begins to develop after the earlier of the two landmarks that we discuss in this section, which is the appearance of what Lewis (1957, pp. 15–16) called “comfort-sounds” as distinct from the “discomfort-cries” that are the infant’s first vocalizations. Although Lewis’s term focuses on the caretaker’s interpretation of the function, his description of the difference in sound quality matches the definition that Oller and Eilers (1988) give to infant productions that they call “fully resonant nuclei” (also termed “syllabic sounds” by Hsu et al., 2000). These are a type of non-distress vocalization with a measurable fundamental frequency throughout most of their duration that are perceived as fully oral vowel-like sounds, as distinct from the markedly nasal quality of the very young infant’s cries and grunts. As Kent and Murray (1982, p. 353) and others have pointed out, the infant’s ability to produce such fully resonant vowel-like sounds is contingent on a reshaping of the back of the vocal tract that lowers the epiglottis to allow the larynx to disengage from the nasopharyngeal passage so that the oral cavity can act as the primary supralaryngeal filter that shapes the source spectrum of the phonation gesture. Measurements from longitudinal and cross-sectional collections of pediatric X-rays and MRIs suggest that this initial descent of the larynx happens around the end of the second month, although there is further lowering of the larynx over the following 5 to 6 years before the length of the child’s pharynx reaches the length of the oral cavity, as in the adult female (Sasaki et al., 1977; Lieberman et al., 2001; Vorperian et al., 2005; Thom et al., 2006; Barbier et al., 2012). That is, while this initial descent of the larynx is the first step in the progression away from the safer (less prone to choking) configuration of the infant chimpanzee vocal tract, typically-developing children will have been talking for years before they have the same 1:1 pharynx length to oral cavity length ratio that de Boer (2010) describes as optimal (producing the largest range of variation in F2 values relative to the range of variation in F1 values).

The anatomical or physiological significance of this initial descent of the larynx at about 2 months, then, is not that the infant’s vocal tract now matches the mother’s, since it does not and will not do so for several years. Rather, the significance of this landmark is that an exhalation-phase phonation gesture can now be coordinated with varying oral gestures, so that the infant can begin to practice posturing the tongue and lips to produce contrasting vowel timbres in combination with the various voice qualities and melodies that the infant already has been learning to control by practicing maneuvers of the respiratory-laryngeal system in the production of other early vocalizations such as cries, whimpers, and grunts (see, e.g., Boliek et al., 1996).

The effect of this practice is evident in two cross-sectional studies of English-learning infants including infants who were older than 2 months but younger than the onset of canonical babble. In the first study, Kent and

Murray (1982) measured F1 and F2 values in fully resonant nuclei produced by seven 3-month-old infants, six 6-month-old infants, and seven 9-month-old infants, recorded during play interactions with their mothers and the experimenters. They found a dramatic expansion in the F2 dimension of the vowel formant space between the youngest group and the 6-month-old group, as well as a further expansion in the F2 dimension between the 6- and 9-month-old infants, with the direction of the expansion indicating a greater and greater control of the front ([i] and [æ]) corners of the vowel quadrangle. In the second study, Kuhl and Meltzoff (1996) measured F1 and F2 values in vocalizations produced by 24 infants in each of three age groups (12-, 16-, and 20-week-olds) in response to audio-visual stimuli of a woman's voice and face saying sequences of [a], [i], or [u] vowels. Measured vocalizations "had to be produced on an exhalatory breath with a visibly open mouth, be relatively steady state, and have an audible voice pitch" (Kuhl and Meltzoff, 1996, p. 2428). Kuhl and Meltzoff found the same expansion in the F2 dimension into the [i] corner of the vowel space. It is important to note that this expansion of the range of F2 values produced is the opposite of what would be predicted from the overall lengthening of the vocal tract alone. Therefore, it indicates an increasing control over lingual and labial gestures independent of the jaw in the several months before the onset of canonical babbling. That is, this expansion pre-figures the further age-related shifts in the center of the vowel formant space that Rvachew et al. (2006) show in their cross-sectional sectional study of the fully-resonant nuclei and the vocalic portion of canonical syllables produced by 43 infants ranging in age from 10 to 18 months, with up to 60 such vowels analyzed for each infant from a recording of a 30-minute session of play with the mother.

The Rvachew et al. (2006) study included infants growing up in French-speaking homes as well as infants growing up in English-speaking homes, and the age-related shifts are in different dimensions for the two groups. That is, for the 24 English-learning infants, the shift in the center of the vowel space is primarily in the F2 dimension, as might be expected from the dimension of expansion in the two studies that examined fully resonant nuclei produced by infants before the onset of canonical babble. By contrast, for the 27 French-learning infants, the shift in the center of the vowel space is primarily in the F1 dimension. Moreover, there are cross-language differences in the distribution of formant values even for the youngest infants in the Rvachew et al. (2006) study, which agree with the differences between the F1-F2 spaces for the five English-learning versus five French-learning infants in the study by de Boysson–Bardies et al. (1989) of vowels produced in canonical babble by 10-month-old infants growing up in English-, French-, Arabic-, or Cantonese-speaking homes. This cross-language difference is in keeping with the differences in the language-specific "articulatory setting" documented by Wilson (2006). It suggests that already by the onset of canonical babble, infants have been exercising timbre-producing labial and lingual postures that are important for the language-specific vowel space.

In other words, as we noted earlier, the deictic game is not an appropriate choice for modeling the interactions between social agents that are relevant for the emergence of vowel inventories. It is not appropriate because the evidence is that, in at least five different cultures, infants with normal hearing begin to produce vowel-like vocalizations that are appropriate for the vowel phoneme inventory of the ambient language months before there is any evidence of the attentional capacity to engage in triadic joint attention to an external referent. To simulate the emergence of this language-specific vowel space in ontogeny, then, it is important to be able to model the kinds of social interaction that are available to infants between the first and the second of the two developmental landmarks discussed in this section.

In the literature on language socialization routines across cultures, the discussion of culture-typical responses to infants' pre-verbal vocalizations does not tend to differentiate between responses to the fully-resonant nuclei produced before 6 months and responses to the more word-like canonical babble produced after 8 months. Therefore, our review of the social significance of the earlier landmark is limited to descriptions from a very small set of cultures, all of which happen to be cultures in which infant-mother dyads interact in face-to-face vocal play. Lewis (1957, p. 31) connects the emergence of vowel-like comfort-sounds with the onset of smiling as a response to the mother's voice, around the end of the second month, and Masataka (1993), Hsu et al. (2001), and others have documented the contingencies between infants' productions of such fully resonant nuclei and infant and mother interactions such as smiling and mutual gaze during the third and fourth months of life. Kent and Murray (1982) and Hsu et al. (2000), among many others, also have noted the melodic variability and complexity of such fully resonant vowel-like sounds, which matches the observed melodic variability and complexity in caregivers' utterances that Papoušek et al. (1991) have associated with interactive contexts such as eliciting the infant's attention, encouraging the infant to vocalize in turn, and contingent rewarding of the infant's imitation of the caretaker's vocalization.

The similarity between this type of dyadic interaction and the imitation game used in modeling studies such as those of de Boer (2000) comes out very clearly in a longitudinal study described by Masataka (2003, pp. 104–123). In this study, face-to-face vocal interactions between ten Japanese mothers and their 3- to 4-month-old infants were recorded in their homes in sets of nine 15-minute sessions spread over three days every two weeks, to make 10 sets of recordings from the time the infant was 8 weeks to when the infant was 28 weeks of age. In the analyses of the recordings, interaction "episodes" were defined as series of vocalizations in turn, with pauses of no more than 300 ms separating the utterances within an episode and pauses of at last 5 minutes between episodes, as in the definition of stimulus and response coo call pairs in Sugiura (1998). The first two utterances were extracted from each episode, distinguishing between mother-initiated interactions (where the infant responded to the mother's episode-initial vocalization) and infant-initiated episodes (where the mother responded to the infant's episode-initial vocalization). Two transcribers classified the mothers' vocalizations using the five vowel categories of Japanese, and then interactions where utterances were classified as /e/ or /o/, which constituted less than 3% of the mothers' utterances, were removed from the analysis. F1 and F2 values were measured in all of the infants' vocalizations. Classifications were also made of the mothers' pitch patterns (using categories such as "falling" versus "rising" versus "bell-shaped") and fundamental frequency measures indicative of these categories were made for the infants' vocalizations. Discriminant analyses predicting the mother's vowel category from the infant's formant values and the mother's pitch pattern category from the infant's fundamental frequency measures in the infant-initiated interactions showed that some mothers were initially imitating the infant's timbre patterns and others were initially imitating the infant's pitch patterns, but in both groups, the maternal imitation rate declined over the 10 weeks of the study. Discriminant analyses predicting the mother's vowel category from the infant's formant values in the mother-initiated interactions at each of the 10 sets of recordings, by contrast, showed that infants were imitating their mothers' vowels more and more across the 10 weeks of the study. Moreover, this relationship was strongest for dyads where the mother initially imitated the infant's timbre pattern rather than the infant's pitch pattern.

4. Summary

The COSMO framework put forward by MDSB is meant to address the issue of the origin of substantive language universals by first establishing a basis of cognitive and communicational principles governing communities of interchangeable prelinguistic agents who engage with each other in vocal interactions, and then demonstrating that such universals can emerge from series of vocal interactions between the agents that are constrained by the established cognitive and communicational principles. Specifically, MDSB assume the following principles regarding agents' cognitive and communicational capacities: *adequacy* – the agents' signals are easy for agents to both perceive and produce, *parity* – there exists a correspondence between the motor repertoire of an agent acting as a speaker and the perceptual repertoire of an agent acting as a listener, and *reference* – each of the agents knows the relationship between their motor and perceptual repertoires (and those of the other agents) and objects in the external world. MDSB's implementation choices for modeling these principles yields agents with developmentally advanced perception-production capabilities whose vocal interactions are restricted to triadic deictic games involving two interchangeable agents and their joint attention to a shared reference object. Below we summarize our arguments and conclusions concerning MDSB's assumptions about language universals and their potential explanations, and some of the consequences of the choices they have made in implementing the COSMO framework as a means for investigating language universals.

The literature reviewed in section 2 leads us to conclude that the COSMO model's VLAM-based reference frames that simulate the vocal tract of a single adult female talker are incommensurate (to varying degrees) with the "phonological" reference frame derived from IPA symbols and sound classification criteria used in creating the UPSID database. Since MDSB's evaluation of the COSMO model relies on a direct mapping from representations in the former frame to those in the latter in order to carry out quantitative comparison of the model's output to cross-language statistics derived from the set of languages described in the UPSID database, we conclude that the evaluation metric is qualitatively flawed. Moreover, we conclude that the COSMO framework is, in its present form, unsuitable for modeling crucially important examples of the emergence of phonological systems at the time-scale of ontogeny. For example, phonological emergence during early infancy involves, at the very least, two agents, say a mother and infant, who are clearly not interchangeable (the infant is prelinguistic while the mother has language, the mother's perception-production capabilities are far more developmentally advanced than those of the infant, the vocal tracts of the agents differs substantially, etc.), and whose initial vocal interactions are more properly conceived of as a dyadic imitation game between the agents that facilitates the infant's formation of a culture-specific phonological system rather than as a triadic deictic game involving the agents' joint attention to a shared reference object. However, we note that the COSMO framework may still be potentially suitable for modeling certain limited sets of phenomena occurring at the time-scales of cultural and biological evolution, such as the emergence of phonological systems within newly formed communities of humans who lack language (as in the case of Al-Sayyid Bedouin Sign Language, see Sandler et al., 2005), or the evolution of vocal communication systems with functional reference resulting from long-term biological conditioning (as in the case of animal alarm calls, see, e.g., Evans et al., 1993, Manser et al., 2001; Hollén and Manser, 2007, Seyfarth et al., 1980, etc.) that exhibit a kind of "adaptive dispersion" but lack the extreme signaling diversity unique to human language.

Crucially, the literature reviewed in section 2 also calls into question the assumptions made by MDSB regarding exactly what counts as a language universal, and by extension, what is in need of explanation via conceptual/computational frameworks such as COSMO. Specifically, we claim that the purported universals taken by MDSB as the motivation for developing the COSMO framework are largely performative artifacts that result from more fundamental prosodic aspects of human language that provide the means for its sequential and simultaneous compositionality, and in section 3, we construct an argument in support of this claim. The context of the argument is mostly restricted to the time-scale of ontogeny, and the argument itself centers on two long-recognized developmental landmarks of phonological acquisition: the emergence between 6 and 8 months of "canonical babbling" or "canonical syllable" (see Oller and Eilers, 1988), and the appearance of "comfort-sounds" (see Lewis, 1957), also termed "fully resonant nuclei" (Oller and Eilers, 1988) or "syllabic sounds" (Hsu et al., 2000), following the descent of the larynx at around 2 months. Experimental evidence suggests that while caretakers differ cross-culturally in their interpretations of canonical babbling, it is this simultaneous coordination of rhythmic mandibular motions with different postures of the tongue and lips and an exhalation-phase phonation gesture that nevertheless provides the ontogenetic basis for the generalization that across all spoken languages words have a productive internal serial structure. Moreover, evidence concerning infants' comfort sounds suggests that the descent of the larynx provides the latitude needed to combine exhalation-phase phonation gestures with various different oral gestures so that infants may begin to attune their gestural combinations to the language-specific vowel spaces they encounter during affiliative interactions with caretakers prior to the onset of canonical babbling and the emergence of clearly referential vocalizations. If so, it follows that the COSMO model's deictic games based on reference are inappropriate for modeling phonological emergence during early infancy. We also note that, from a comparative perspective, a number of other primate species also exhibit the use of signal-internal rhythmicity in their affiliative exchanges and at least some of these species also exhibit the combination of laryngeal and oral gestures in this rhythmic structuring of vocal signals in these affiliative exchanges.

This commentary taken as a whole suggests that a renewed attention to the experimental literature on affiliative exchanges between infants and their caretakers, as well as the very broad range of affiliative exchanges between primate conspecifics, may serve as a fruitful basis for the development of conceptual and computational frameworks for understanding the emergence of phonological universals. In other work (see, e.g., Plummer, 2014; Plummer and Beckman, submitted), we describe a conceptual framework and family of computational frameworks that we have been developing toward that goal.

Acknowledgements

Work supported by NIH grant R01 DC02932 to Jan Edwards, by NSF grant BCS0729306 to Mary E. Beckman, by NSF grant BCS0729277 to Benjamin Munson, and by NSF grant CNS-1305319 to Eric Fosler-Lussier. We thank the editors and reviewers, as well as Pat Reidy and Becca Morley, for their comments which helped us greatly in seeing how to clarify our exposition.

References

Arcadi, A.C., 1996. Phrase structure of wild chimpanzee pant hoots: Patterns of production and interpopulation variability. American Journal of Primatology 39, 159–178. doi:10.1002/(SICI)1098-2345(1996)39:3<159::AID-AJP2>3.0.CO;2-Y.

Arcadi, A.C., 2000. Vocal responsiveness in male wild chimpanzees: Implications for the evolution of language. Journal of Human Evolution 39, 205–223. doi:10.1006/jhev.2000.0415.

Barbier, G., Boë, L.J., Captier, G., 2012. La croissance du conduit vocal du foetus á l'adulte: Une étude longitudinale. Biométrie Humaine et Anthropologie 30, 11–22.

Barbier, G., Boë, L.J., Captier, G., Laboissière, R., 2015. Human vocal tract growth: A longitudinal study of the development of various anatomical structures, in: Proceedings of INTERSPEECH 2015.

Baronchelli, A., Chater, N., Pastor-Satorras, R., Christiansen, M.H., 2012. The biological origin of linguistic diversity. PLoS ONE 7, e48029. doi:10.1371/ journal.pone.0048029.

Beckman, M.E., Edwards, J., 2010. Generalizing over lexicons to predict consonant mastery. Laboratory Phonology 1, 319–343.

Bell, A., 1978. Syllabic consonants, in: Greenberg, J.H., Ferguson, C.A., Moravcsik, E.A. (Eds.), Universals of human language. volume 2: Phonology, pp. 153–201.

Blount, B.G., 1972. Aspects of Luo socialization. Language in Society 1, 235–248. doi:10.1017/S0047404500000518.

Boë, L., Maeda, S., 1998. Modélisation de la croissance du conduit vocal. Espace vocalique des nouveaux–nés et des adultes. Conséquences pour l'ontegenèse et la phylogenèse, in: Journées d'Études Linguistiques: "La Voyelle dans Tous ces États". Nantes, France, pp. 98–105.

de Boer, B., 2000. Self–organization in vowel systems. Journal of Phonetics 28, 441–465. doi:10.006/jpho.2000.0125.

de Boer, B., 2010. Investigating the acoustic effect of the descended larynx with articulatory models. Journal of Phonetics 38, 679–686. doi:10.1016/j.wocn.2010.10.003.

Boliek, C.A., Hixon, T.J., Watson, P.J., Morgan, W.J., 1996. Vocalization and breathing during the first year of life. Journal of Voice 10, 1–22. doi:10.1016/S0892-1997(96)80015-4.

de Boysson–Bardies, B., Halle, P., Sagart, L., Durand, C., 1989. A crosslinguistic investigation of vowel formants in babbling. Journal of Child Language 16, 1–17. doi:10.1017/S0305000900013404.

Braem, P.B., 1999. Rhythmic temporal patterns in the signing of deaf early and late learners of Swiss German Sign Language. Language and Speech 42, 177–208. doi:10.1177/00238309990420020301.

Brockelman, W.Y., Reichard, U., Treesucon, U., Raemaekers, J.J., 1998. Dispersal, pair formation and social structure in gibbons (Hylobates lar). Behavioral Ecology and Sociobiology 42, 329–339. doi:10.1007/s002650050445.

Chater, N., Christiansen, M.H., 2010. Language acquisition meets language evolution. Cognitive Science 34, 1131–1157. doi:10.1111/j.1551-6709.2009.01049.x.

Chung, H., Kong, E.J., Edwards, J., Weismer, G., Fourakis, M., Hwang, Y., 2012. Cross-linguistic studies of children's and adults' vowel spaces. Journal of the Acoustical Society of America 131, 442–454. doi:10.1121/1.3651823.

Clarke, E., Reichard, U.H., Zuberbühler, K., 2006. The syntax and meaning of wild gibbon songs. PLoS One 1, e73. doi:10.1371/ journal.pone.0000073.

Cohen, E., 2012. The evolution of tag-based cooperation in humans: The case for accent. Current Anthropology 53, 588–616. doi:10.1086/667654.

Crelin, E.S., 1969. Anatomy of the Newborn: An Atlas. Lea and Febiger, Philadelphia.

Crelin, E.S., 1987. The human vocal tract: Anatomy, function, development, and evolution. Vantage Press, New York.

Crockford, C., Herbinger, I., Vigilant, L., Boesch, C., 2004. Wild chimpanzees produce group-specific calls: A case for vocal learning? Ethology 110, 221–243. doi:10.1111/j.1439-0310.2004.00968.x.

Disner, S.F., 1983. Vowel quality: The relation between universals and language-specific factors. UCLA Working Papers in Phonetics 58, 1–158.

Elowson, A.M., Snowdon, C.T., Lazaro-Perea, C., 1998. Infant 'babbling' in a nonhuman primate: Complex vocal sequences with repeated call types. Behaviour 135, 643–664. URL: http://www.jstor.org/stable/4535550, doi:10.1163/156853998792897905.

Evans, C.S., Evans, L., Marler, P., 1993. On the meaning of alarm calls: Functional reference in an avian vocal system. Animal Behaviour 46, 23–38. doi:10.1006/anbe.1993.1158.

Fedurek, P., Schel, A.M., Slocombe, K.E., 2013. The acoustic structure of chimpanzee pant-hooting facilitates chorusing. Behavioral Ecology and Sociobiology 67, 1781–1789. doi:10.1007/s00265-013-1585-7.

Ferrari, P.F., Paukner, A., Ionica, C., Suomi, S.J., 2009. Reciprocal face-to-face communication between rhesus macaque mothers and their newborn infants. Current Biology 19, 1768–1772. doi:10.1016/j.cub.2009.08.055.

Ferrari, P.F., Visalberghi, E., Paukner, A., Fogassi, L., Ruggiero, A., Suomi, S.J., 2006. Neonatal imitation in rhesus macaques. PLoS Biology 4, e302. doi:10.1371/journal.pbio.0040302.

Geissmann, T., 1999. Duet songs of the Siamang, Hylobates syndactylus: II. Testing the pair-bonding hypothesis during a partner exchange. Behaviour 136, 1005–1039. URL: http://www.jstor.org/stable/4535656, doi:10.1163/156853999501694.

Ghazanfar, A.A., Takahashi, D.Y., 2014. Facial expressions and the evolution of the speech rhythm. Journal of Cognitive Neuroscience 26, 1196–1207. doi:10.1162/jocn_a_00575.

Ghazanfar, A.A., Takahashi, D.Y., Mathur, N., Fitch, W.T., 2012. Cineradiography of monkey lip-smacking reveals putative precursors of speech dynamics. Current Biology 22, 1176–1182. doi:10.1016/j.cub.2012.04.055.

Goodall, J., 1986. The chimpanzees of Gombe: Patterns of behavior. Cambridge University Press, Cambridge, MA.

Greenberg, J.H., 1965. Some generalizations concerning initial and final consonant sequences. Linguistics 18, 5–34. doi:10.1515/ling.1965.3.18.5.

Griebel, U., Oller, D.K., 2008. Evolutionary forces favoring contextual flexibility: The role of deception and protean behavior, in: Oller, D.K., Griebel, U. (Eds.), Evolution of communicative flexibility: Complexity, creativity, and adaptability in human and animal communication. MIT Press, Cambridge, MA. chapter 2, pp. 9–40.

Haimoff, E.H., 1986. Convergence in the duetting of monogamous old world primates. Journal of Human Evolution 15, 51–59. doi:10.1016/S0047-2484(86)80065-3.

Hinde, R.A., Rowell, T.E., 1962. Communication by posture and facial expressions in the rhesus monkey (Macaca mulatta). Proceedings of the Zoological Society, London 138, 1–21. doi:10.1111/j.1469-7998.1962.tb05684.x.

Hockett, C.F., Ascher, R., 1964. The human revolution. Current Anthropology 5, 135–147. URL: http://www.jstor.org/stable/2740176.

Hollén, L.I., Manser, M.B., 2007. Motivation before meaning: Motivational information encoded in meerkat alarm calls develops earlier than referential information. The American Naturalist 169, 758–767. URL: http://www.jstor.org/stable/10.1086/516719, doi:10.1086/516719.

Hsu, H.C., Fogel, A., Cooper, R.B., 2000. Infant vocal development during the first 6 months: Speech quality and melodic complexity. Infant and Child Development 9, 1–16. doi:10.1002/(SICI)1522-7219(200003)9:1<1::AID-ICD210>3.0.CO;2-V.

Hsu, H.C., Fogel, A., Messinger, D.S., 2001. Infant non-distress vocalization during mother-infant face-to-face interaction: Factors associated with quantitative and qualitative differences. Infant Behavior and Development 24, 107–128. doi:10.1016/S0163-6383(01)00061-3.

Ishihara, H., Yoshikawa, Y., Miura, K., Asada, M., 2009. How caregiver's anticipation shapes infant's vowel through mutual imitation. Autonomous Mental Development, IEEE Transactions on 1, 217–225. doi:10.1109/TAMD.2009.2038988.

Ishizuka, K., Mugitani, R., Kato, H., Amano, S., 2007. Longitudinal developmental changes in spectral peaks of vowels produced by Japanese infants. Journal of the Acoustical Society of America 121, 2272–2282. doi:10.1121/1.2535806.

Jakobson, R., Gunnar, F., Halle, M., 1951/1969. Preliminaries to speech analysis: The distinctive features and their correlates. MIT Press, Cambridge, MA.

Kent, R.D., 1984. The psychobiology of speech development: Coemergence of language and a movement system. American Journal of Physiology 246, R888–R894.

Kent, R.D., Murray, A.D., 1982. Acoustic features of infant vocalic utterances at 3, 6, and 9 months. Journal of the Acoustical Society of America 72, 353–365. doi:10.1121/1.388089.

Kirby, S., Hurford, J.R., 2002. The emergence of linguistic structure: An overview of the iterated learning model, in: Cangelosi, A., Parisi, D. (Eds.), Simulating the evolution of language. Springer, London, pp. 121–147. doi:10.1007/978-1-4471-0663-0_6.

Koda, H., Lemasson, A., Oyakawa, C., Rizaldi, Pamungkas, J., Masataka, N., 2013. Possible role of mother-daughter vocal interactions on the development of species-specific song in gibbons. PLoS ONE 8, e71432. doi:10.1371/journal.pone.0071432.

Kuhl, P.K., Meltzoff, A.N., 1996. Infant vocalizations in response to speech: Vocal imitation and developmental change. Journal of the Acoustical Society of America 100, 2425–2438. doi:10.1121/1.417951.

Kulick, D., Schiefflin, B.B., 2004. Language socialization, in: Duranti, A. (Ed.), A companion to linguistic anthropology. Blackwell, Malden, MA. chapter 15, pp. 349–368. doi:10.1002/9780470996522.ch15.

Ladd, D.R., 2014. Phonetics in phonology, in: Simultaneous Structure in Phonology. Oxford University Press, Oxford. chapter 2, pp. 29–55.

Ladefoged, P., 1983. The limits of biological explanations in phonetics. UCLA Working Papers in Phonetics 57, 1–10.

Ladefoged, P., Bhaskararao, P., 1983. Non-quantal aspects of consonant production: A study of retroflex consonants. Journal of Phonetics 11, 291–302.

Levitt, A.G., Wang, Q., 1991. Evidence for language-specific rhythmic influences in the reduplicative babbling of French- and English-learning infants. Language and Speech 34, 235–249.

Lewis, M.M., 1957. How Children Learn to Speak. Harrap, London.

Lieberman, D.E., McCarthy, R.C., Hiiemae, K.M., Palmer, J.B., 2001. Ontogeny of postnatal hyoid and larynx descent in humans. Archives of Oral Biology 46, 117–128. doi:10.1016/S0003-9969(00)00108-4.

Liljencrants, L., Lindblom, B., 1972. Numerical simulations of vowel quality systems: The role of perceptual contrast. Language 48, 839–862. doi:10.2307/411991.

Lindblom, B., MacNeilage, P.F., Studdert-Kennedy, M., 1984. Self-organizing processes and the explanation of phonological universals, in: Butterworth, B., Comrie, B., Dahl, Ö. (Eds.), Explanations for language universals. Mouton, Berlin, pp. 181–204.

Macken, M.A., Ferguson, C.A., 1981. Phonological universals in language acquisition. Annals of the New York Academy of Sciences 379, 110–129. doi:10.1111/j.1749-6632.1981.tb42002.x.

MacNeilage, P., Davis, B.L., 2000. On the origin of internal structure of word forms. Science 288, 527–531. doi:10.1126/science.288.5465.527.

MacNeilage, P.F., Davis, B.L., Matyear, C.L., 1997. Babbling and first words: Phonetic similarities and differences. Speech Communication 22, 269–277. doi:10.1016/S0167-6393(97)00022-8.

Maddieson, I., 1984. Patterns of Sounds. Cambridge University Press, Cambridge. doi:10.1017/CBO9780511753459.

Maddieson, I., 1991. Testing the universality of phonological generalizations with a phonetically specified segment database: Results and limitations. Phonetica 48, 193–206. doi:10.1159/000261884.

Manser, M.B., Bell, M.B., Fletcher, L.B., 2001. The information that receivers extract from alarm calls in suricates. Proceedings of the Royal Society of London, Series B 268, 2485–2491. doi:10.1098/rspb.2001.1772.

Martinet, A., 1949. La double articulation linguistique. Travaux du Cercle Linguistique de Copenhague 5, 30–37.

Masataka, N., 1993. Effects of contingent and noncontingent maternal stimulation on the vocal behaviour of three– and four–month–old Japanese infants. Journal of Child Language 20, 303–312. doi:10.1017/S0305000900008291.

Masataka, N., 2003. The Onset of Language. Cambridge University Press, Cambridge, UK.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., Amiel-Tison, C., 1988. A precursor of language acquisition in young infants. Cognition 29, 143–178. doi:10.1016/0010-0277(88)90035-2.

Meir, I., Sandler, W., Padden, C., Aranoff, M., 2010. Emerging sign languages, in: Marschark, M., Spencer, P.E. (Eds.), The Oxford handbook of deaf studies, language, and education. Oxford University Press, Oxford. chapter 18, pp. 267–280. doi:10.1093/oxfordhb/9780195390032.013.0018.

Mitani, J.C., 1985. Gibbon song duets and intergroup spacing. Behaviour 92, 59–96. doi:10.1163/156853985X00389.

Mitani, J.C., 2009. Cooperation and competition in chimpanzees: Current understanding and future challenges. Evolutionary Anthropology 18, 215–227. doi:10.1002/evan.20229.

Mitani, J.C., Gros-Louis, J., 1998. Chorusing and call convergence in chimpanzees: Tests of three hypotheses. Behaviour 135, 1041–1064. URL: http://www.jstor.org/stable/4535578.

Miura, K., Yoshikawa, Y., Asada, M., 2012. Vowel acquisition based on an auto–mirroring bias with a less imitative caregiver. Advanced Robotics 26, 23–44. doi:10.1163/016918611X607347.

Morrill, R.J., Paukner, A., Ferrari, P.F., Ghazanfar, A.A., 2012. Monkey lipsmacking develops like the human speech rhythm. Developmental Science 15, 557–568. doi:10.1111/j.1467-7687.2012.01149.x.

Moulin-Frier, C., Diard, J., Schwartz, J.L., Bessière, P., 2016. Cosmo ("communicating about objects using sensory-motor operations"): a bayesian modeling framework for studying speech communication and the emergence of phonological systems. Journal of Phonetics tbd.

Nartey, J.N.A., 1979. A study in phonetic universals: Especially concerning fricatives and stops. Ph.D. thesis. University of California, Los Angeles.

Nathani, S., Oller, D.K., Neal, A.R., 2007. On the robustness of vocal development: An examination of infants with moderate-to-severe hearing loss and additional risk factors. Journal of Speech, Language, and Hearing Research 50, 1425–1444. doi:10.1044/1092-4388(2007/099).

Nazzi, T., Bertoncini, J., Mehler, J., 1998. Language discrimination by newborns: Toward an understanding of the role of rhythm. Journal of Experimental Psychology: Human Perception and Performance 24, 756–766. doi:10.1037/0096-1523.24.3.756.

Ochs, E., Schieffelin, B.B., 1982. Language acquisition and socialization: Three developmental stories and their implications. Number 105 in Sociolinguistic Working Paper, Southwest Educational Development Laboratory, Austin, Texas.

Ohala, J.J., 1983. The origin of sound patterns in vocal tract constraints, in: MacNeilage, P.F. (Ed.), The production of speech. Springer-Verlag, New York. chapter 9, pp. 189–216. doi:10.1007/978-1-4613-8202-7_9.

Oller, D.K., 1980. The emergence of the sounds of speech in infancy, in: Yeni-Komshian, G., Kavanagh, J., Ferguson, C. (Eds.), Child phonology, Volume 1, Production. Academic Press, New York, pp. 93–112.

Oller, D.K., Eilers, R.E., 1988. The role of audition in infant babbling. Child Development 59, 441–449. doi:10.2307/1130323.

Papoušek, M., Papoušek, H., Symmes, D., 1991. The meanings of melodies in motherese in tone and stress languages. Infant Behavior and Development 14, 415–440. doi:10.1016/0163-6383(91)90031-M.

Parr, L.A., Cohen, M., de Waal, F., 2005. Influence of social context on the use of blended and graded facial displays in chimpanzees. International Journal of Primatology 26, 73–103. doi:10.1007/s10764-005-0724-z.

Plummer, A.R., 2014. The acquisition of vowel normalization: Theory and computational framework. Ph.D. thesis. The Ohio State University.

Plummer, A.R., Beckman, M.E., submitted. Developing conceptual and computational frameworks for modeling the earliest phonological abstraction in infant-caretaker vocal imitation. Journal of Phonetics .

Rasilo, H., Räsänen, O., Laine, U.K., 2013. Feedback and imitation by a caregiver guides a virtual infant to learn native phonemes and the skill of speech inversion. Speech Communication 55, 909–931. doi:10.1016/j.specom.2013.05.002.

Richman, A.L., Miller, P.M., LeVine, R.A., 1992. Cultural and educational variations in maternal responsiveness. Developmental Psychology 28, 614–621. doi:10.1037/0012-1649.28.4.614.

Ross, M.D., Owren, M.J., Zimmermann, E., 2009. Reconstructing the evolution of laughter in great apes and humans. Current Biology 19, 1106–1111. doi:10.1016/j.cub.2009.05.028.

Rvachew, S., Alhaidary, A., Mattock, K., Polka, L., 2008. Emergence of the corner vowels in the babble produced by infants exposed to Canadian English or Canadian French. Journal of Phonetics 36, 564–577. doi:10.1016/j.wocn.2008.02.001.

Rvachew, S., Mattock, K., Polka, L., Ménard, L., 2006. Developmental and cross–linguistic variation in the infant vowel space: The case of Canadian English and Canadian French. Journal of the Acoustical Society of America 120, 2250–2259. doi:10.1121/1.2266460.

Sandler, W., Meir, I., Padden, C., Aronoff, M., 2005. The emergence of grammar: Systematic structure in a new language. Proceedings of the National Academy of Sciences of the United States of America 102, 2661–2665. doi:10.1073/pnas.0405448102.

Sasaki, C.T., Levine, P.A., Laitman, J.T., Crelin, E.S., 1977. Postnasal descent of the epiglottis in man: A preliminary report. Archives of Otolaryngology – Head & Neck Surgery 103, 169–171.

Senghas, A., Kita, S., Özyürek, A., 2004. Children creating core properties of language: Evidence from an emerging sign language in Nicaragua. Science 305, 1779–1782. doi:10.1126/science.1100199.

Seyfarth, R.M., Cheney, D.L., 2003. Signalers and receivers in animal communication. Annual Review of Psychology 54, 145–173. doi:10.1146/annurev.psych.54.101601.145121.

Seyfarth, R.M., Cheney, D.L., Marler, P., 1980. Vervet monkey alarm calls: Semantic communication in a free-ranging primate. Animal Behaviour 28, 1070–1094. doi:10.1016/S0003-3472(80)80097-2.

Snow, C.P., 1977. The development of conversation between mothers and babies. Journal of Child Language 4, 1–22. doi:10.1017/S0305000900000453.

Snowdon, C.T., Cronin, K.A., 2007. Cooperative breeders do cooperate. Behavioural Processes 76, 138–141. doi:10.1016/j.beproc.2007.01.016.

Snowdon, C.T., Elowson, A.M., 1999. Pygmy marmosets modify call structure when paired. Ethology 105, 893–908. doi:10.1046/j.1439-0310.1999.00483.x.

Stevens, K.N., 1972. The quantal nature of speech: Evidence from articulatory–acoustic data, in: David, E.E., Denes, P.B. (Eds.), Human communication: A unified view. McGraw–Hill, pp. 51–66.

Stevens, K.N., 1989. On the quantal nature of speech. Journal of Phonetics 17, 3–45.

Stoel-Gammon, C., Williams, K., Buder, E., 1994. Cross-language differences in phonological acquisition: Swedish and American /t/. Phonetica 51, 146–158. doi:10.1159/000261966.

Sugiura, H., 1998. Matching of acoustic features during the vocal exchange of coo calls by Japanese macaques. Animal Behavior 55, 673–687. doi:10.1006/anbe.1997.0602.

Sugiura, H., 2007. Effects of proximity and behavioral context on acoustic variation in the coo calls of Japanese macaques. American Journal of Primatology 69, 1412–1424. doi:10.1002/ajp.20447.

Tanaka, T., Sugiura, H., Masataka, N., 2006. Cross-sectional and longitudinal studies of the development of group differences in acoustic features of coo calls in two groups of Japanese macaques. Ethology 112, 7–21. doi:10.1111/j.1439-0310.2006.01103.x.

Thom, S.A., Hoit, J.D., Hixon, T.J., Smith, A.E., 2006. Velopharyngeal function during vocalization in infants. Cleft Palate-Craniofacial Journal 43, 539–546. doi:10.1597/05-113.

Vihman, M.M., 1993. Variable paths to early word production. Journal of Phonetics 21, 61–82.

Vihman, M.M., 2014. Phonological development: The first two years. 2nd ed., Wiley–Blackwell, Malden, MA.

Vogt, P., 2005. On the acquisition and evolution of compositional languages: Sparse input and the productive creativity of children. Adaptive Behavior 13, 325–346. doi:10.1177/105971230501300403.

Vorperian, H.K., Kent, R.D., Lindstrom, M.J., Kalina, C.M., Gentry, L.R., Yandell, B.S., 2005. Development of vocal tract length during early childhood: A magnetic resonance imaging study. Journal of the Acoustical Society of America 117, 338–350. doi:10.1121/1.1835958.

Vorperian, H.K., Wang, S., Chung, M.K., Schimek, E.M., Durtschi, R.B., Kent, R.D., Ziegert, A.J., Gentry, L.R., 2009. Anatomic development of the oral and pharyngeal portions of the vocal tract: An imaging study. Journal of the Acoustical Society of America 125, 1666–1678. doi:10.1121/1.3075589.

Wilson, I.L., 2006. Articulatory settings of French and English monolingual and bilingual speakers. Ph.D. thesis. University of British Columbia.