Empirical approaches for Investigating the Origins of Structure in Speech

Hannah Little[1,2], Heikki Rasilo[1,3], Sabine van der Ham[1], and Kerem Eryılmaz[1]

[1] Artificial Intelligence Lab, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

[2] Department of Language and Cognition, Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD, Nijmegen, The Netherlands

[3] Department of Signal Processing and Acoustics, Aalto University, Otakaari 5 A, 02150 Espoo, Finland

Abstract

In language evolution research, the use of computational and experimental methods to investigate the emergence of structure in language is exploding. In this review, we look exclusively at work exploring the emergence of structure in speech, on both a categorical level (what drives the emergence of an inventory of individual speech sounds), and a combinatorial level (how these individual speech sounds emerge and are reused as part of larger structures). We show that computational and experimental methods for investigating population-level processes can be effectively used to explore and measure the effects of learning, communication and transmission on the emergence of structure in speech. We also look at work on child language acquisition as a tool for generating and validating hypotheses for the emergence of speech categories. Further, we review the effects of noise, iconicity and production effects.

Keywords: evolution of speech; combinatorial structure; phonetic learning; artificial language experiments

1. Introduction

Structure in speech can generally be understood on two levels, category structure (i.e. the structure of speech sound inventories), and combinatorial structure (i.e. how those sounds are combined into meaningful units). How these levels of structure, as we know them today, emerged in the language of our ancestors cannot be observed directly, because all spoken human languages already have fully established sound systems. Accordingly, we need to instead look to more indirect forms of evidence, which are reviewed in this article.

Language evolved as part of a complex adaptive system whereby ontogeny (individual linguistic development), glossogeny (population level cultural evolution of language) and phylogeny (biological evolution of language) interact (Kirby, 2002). The majority of work on the evolution of speech has its focus on the biological evolution of the vocal tract (Fitch, 2000). However, this article has its focus on ontogeny and glossogeny. We review recent empirical methods, primarily from language evolution research, outlining how they can specifically be used to look at speech, rather than language in a broader sense. The article is therefore aimed at those interested in language evolution who wish to co-opt existing empirical methods to look at speech, or those interested in speech who wish to gain knowledge on how we might study its evolution from a cultural rather than a biological perspective. Further, we offer a broad perspective of the work that has already done highlight some gaps of where more work needs to be done.

We will cover computational modelling, ranging from cognitively plausible models of language acquisition, to much more abstract models of how language behaves at a population level. We also cover some empirical findings of speech category learning in infants at the ontogenetic level, as this is relevant to how categories might have formed as a result of the transmission of language. Further, we cover artificial language learning (ALL) experiments looking at how acquisition extrapolates to the population level using iterated learning paradigms, and experiments looking at how interaction might affect cultural evolution using

social coordination experiments. Studies investigating population level processes using ALL experiments were originally used in language evolution research to look at the emergence of compositional structure (e.g. Kirby, Cornish, & Smith, 2008). Compositional structure is the level of structure where meaningful units (e.g. morphemes and words) are combined to make bigger meaningful units, which is distinct from combinatorial structure where meaningless units (e.g. phonemes) combine to make meaningful units (de Boer, Sandler, & Kirby, 2012). When experimentally investigating compositional structure, studies tend to have a starting point of discretised units, most often artificial syllables constructed from existing written characters of the Roman alphabet (e.g. Kirby et al., 2008; Silvey, Kirby, & Smith, 2013). However, in order to investigate the emergence of categorical or combinatorial structure, studies need to use signals created using a continuous signal space. Experiments adopting continuous signal spaces have been adopted mostly in an attempt to validate the assumptions and claims of computational work.

We will first review computational, experimental and developmental approaches that investigate different hypotheses related to the emergence of speech sound inventories. We will then move on to look at how combinatorial structure in speech emerged in experimental and computational studies, and how we might understand the structure we see emerging in these experiments. Throughout, we will be focussing on how learning and communication affect speech structure, as well as the effects of perception and production.

2. Emergence of Categories in Speech

This section reviews work that investigates how (phonetic) category structure emerges. We break down this work into two main sections. The first looks at how structure emerges at the level of a population, both as a result of vertical transmission (learning) and horizontal transmission (communication). This work aims at explaining the emergence of speech categories through interaction with other individuals. We then go on to look at category

learning at the ontogenetic level by reviewing work on child language acquisition. This work focuses not as much on the methods used, but on language acquisition studies as a tool for generating or validating hypotheses for why our phonetic systems look the way they do. It is not our intention here to make the claim that ontogeny is recapitulating the emergence of categories, but it can give us a big incite into what's important for speech categories at an individual level. Further, language is primarily transmitted to new speakers by the infant speech acquisition process and learnability is an important factor biasing languages to develop into a more easily learnable form (see Oudeyer, 2005; Saffran, 2003).

2.1 Category emergence at the population level

Population dynamics may affect the phonemic inventory of a language as language must be used robustly within a population. Language should be robust towards communication in noisy environments, and should also be easily transmittable (learnable) to new language speakers. First, this section briefly discusses computational simulations that investigate several factors affecting formation of sound systems at the population level, followed by an overview of experimental studies that have investigated categorisation of continuous spaces through interaction, social coordination, and iterated learning across generations.

2.1.1 Computational and corpus studies

One of the first computational models proposing a mechanism for the dispersion of vowel categories was Liljencrants and Lindblom (1972). This model assumes that in order for a language to work effectively inside a population that speech sounds with different roles should be clearly distinguished from each other acoustically (i.e. *the principle of maximal perceptual contrast*). That is, if a language only has 3 vowels, then those vowels should be maximally distinct from one another in the 3 corners of the vowel triangle. Liljencrants and Lindblom (1972) present a mathematical model that predicts the locations of vowel categories in vowel

systems of different sizes. The locations were optimised in a two-dimensional acoustic domain based on acoustic distances between vowels. Their model generated vowel systems that have similarities to real-world vowel systems.

Above, the maximal perceptual contrast aims to aid the *listener* rather than the *speaker* to distinguish between phonemes effectively. Lindblom, Macneilage and Studdert-Kennedy (1984) consider *speaker-based* minimal articulatory effort and articulatory sensory discriminability to predict the emergence of structure in syllabic constructions. The constructions produced by their model resemble natural syllable sets regarding places of articulation, vowel contrasts and phonemic coding (syllable starting and ending points can be shared between different syllables). Carré (1996) also model the emergence of vowel systems as the result of efficient use of an articulatory space. They apply the rule of maximum acoustic contrast to select vowel systems of different sizes. The resulting vowel systems correspond well to the most common human systems.

Further to articulatory models that focus on constraints at an individual level, de Boer (2000) has considered the emergence of vowel systems in population-based simulations, where agents, equipped with articulatory and perceptual systems, interact and adapt their vowel inventories based on success in communication. The simulations show that after a large number of interactions, the shared vowel inventories self-organise into systems that show similarities with human vowel systems. Even though acoustically dispersed vowel systems are preferred, acoustically sub-optimal systems (i.e. in this context, the perceptual distances between the speech sounds are not maximised) can also remain functional and be maintained. Such systems are also found in human languages, for a drastic example Navajo (McDonough, Ladefoged, & George, 1992; Vaux & Samuels, 2015), whose vowel system consists of a non-optimally dispersed set of vowels /i/, /e/, /æ/ and /o/.

Further to articulatory models, a corpus study of Dutch language by ten Bosch (1995) suggests that the amount of acoustic contrast between phonemes is affected by the “functional

load" imposed by the lexicon. That is, acoustic contrast between vowels with a small number of minimal pairs does not need to be as strong as it does between vowels with more minimal pairs. Relatedly, a large corpus study by Wedel, Kaplan and Jackson (2013) revealed that sound pairs that distinguish many words in the lexicon are more likely to be maintained as 2 categories in the sound system compared to those with a low functional load which are more likely to collapse into one category. Sound categories collapse once there is no longer any functional pressure to maintain them, eventually resulting in a shift in the sound system of that language.

There is also a wealth of of work exploring historical change in sound systems caused by things like such as variability in articulation (e.g. Blevins, 2004; Kirby & Sonderegger, 2015). This work may shine further light on why sound categories get distributed in the way they do in modern languages, but is outside of the scope of this review.

Taken together, these studies suggest that articulatory and acoustic pressures, as well as functional pressures from higher level linguistic structures, strongly affect the way a sound system is organised. It is clear that not only articulation and perceptual dispersion of sound systems, but also their use in lexicon and their use on the population level can affect the sound system of a language. The function of speech is to transmit information from the speaker to the listener, meaning suboptimal systems may easily be maintained as long as they remain communicatively efficient and learnable.

2.1.2. Experimental studies

For the emergence of categorical structure in speech specifically, cultural learning experiments are very thin on the ground. However, there is work exploring the emergence of categorical structure in visual semantic spaces. We propose that the mechanisms that account for the emergence of categories in semantic spaces will be similar to those which account for the emergence of structure in speech. Both are subject to the same processes, such as pressures for expressivity (making categories distinct from one another) and learnability. Here, we briefly

review work on semantic categories, before discussing a study that did use more speech-like stimuli. We first outline some findings that show parallels with the computational work discussed above, before discussing how the findings from the work on the emergence of semantic categories can serve as a possible model for future studies on the emergence of (speech) sound categories at a population level.

The experimental work here either takes the form of social coordination or communication games, where participants are asked to communicate a set of meanings by taking it in turns to produce and recognise signals (modelling horizontal transmission), or iterated learning experiments, where one participant's reproductions are used as training or input for the next learner (modelling vertical transmission). Sometimes the paradigms are combined in order to compare the result from these approaches to understand the effect of horizontal and vertical transmission together. One important difference with the computational studies reviewed above is that the experimental studies here have both meaning spaces (visual objects) and signal spaces (usually a typed string consisting of random consonant-vowel syllables), while most computational models of speech category emergence have systems based only on a signal space. An important similarity is that the categorisation happens from a space that is continuous (as opposed to discretised). If the pressures that are present in the computational studies discussed above are purely functional and domain-general, we can expect similar patterns in the experimental studies that investigate emergence of visual categories.

Matthews (2009) conducted an iterated learning experiment where participants learnt labels for meanings that were quadrilateral shapes that differed in a continuous fashion. The first participants learned labels for a randomly structured space whose output was taught to a new participant, and so on for 10 generations. As chains progressed, the language became more structured and easier to learn, as shape categories emerged from the continuous meaning space. Although the different iteration chains revealed some chain-specific variation, the categories in each chain were formed based on the similarity between the shapes. Silvey et al. (2013)

compared participants labelling a two-dimensional continuous visual space alone with participants labelling the space as part of a communication task with a partner. Alone, participants produced optimal categories (learnable and expressive); the space was divided into categories of equal size, and covered contiguous regions in the visual space. In the communication conditions, however, non-optimal category structures emerged. Silvey et al. (2016) have since conducted an iterated learning experiment with the same stimuli as the communication games and found that, over 5 generations of communication games, the pressure for learnability pushed there to be fewer categories that were more contiguous.

The semantic studies described here cannot provide clear evidence for (or against) a clear preference for pushing the categories to the periphery of the continuous space, similar to what we saw in the computational studies reviewed in Section 2.1. However, we do see that the categories are affected by the functional pressures of learnability and expressivity. Future work could lie in a direct comparison between a task with visual stimuli and acoustic stimuli, in order to investigate the domain-generality of the mechanisms involved in these tasks.

There has been one experiment that investigated the mechanisms used in categorising an auditory space. Van der Ham & de Boer (2015) investigated the way that an *acoustic space* can be categorised by individuals, and whether this is influenced by a weak *cognitive bias* for more extreme categories. To do this, they conducted a distributional learning experiment that had no influence from functional pressures. They used a one-dimensional signal space that consisted of an /a:/ sound that varied in pitch in a continuous way. Participants learnt 12 sounds from a distribution of the sounds that had only one skewed peak; either biased towards more high or low pitched sounds. They were then asked to reproduce the 12 sounds they heard from memory. The expectation with functional pressures would be that the reproductions would migrate towards the periphery of the signal space (in the direction of the skew). When there is no functional reason to have extreme categories, if such a system emerged, then a this can be said to be the result of cognitive biases. However, participants' reproductions did not exaggerate the

skewness of the input data, indicating that learning such sound categories is not driven by a bias to shift towards more extreme values of a given range. Instead, participants moved the peak more towards the center of the distribution.There was also no tendency to create categories with a uniform distribution.

2.1.3. Summary

In this section, we have discussed the emergence of category structure at the population level, and how cultural processes interact with the biases of individuals in the population. This work is severely underdeveloped, especially regarding the use of experimental work. This is perhaps due to the difficulties of experimentally investigating that emergence of categories in the perception and production of sound categories with adults that have pre-existing linguistic systems. Perhaps, then, the best source of evidence we have in this area is looking at humans who do not yet have sound categories, i.e. infants.

2.2 Category learning in individuals

As we saw earlier, population behaviour is affected by the biases of individuals in that population. In turn, individual biases are shaped by cognitive and physical traits that are the result of biological evolution. Here, we explore the ontogenetic level of language evolution. Language is primarily transmitted to children, and speech sound categories may be shaped by the details of the learning process. Below we will review experimental and computational work on the emergence of categories at the individual level: infant speech acquisition. Section 2.2.1 focuses on computational and experimental studies investigating statistical learning, a central learning mechanism in language acquisition, while Section 2.2.2 outlines psycholinguistic and computational studies that focus on the role of infant-parent interaction in phonetic acquisition. We conclude that learner-based general associative and motor control learning mechanisms may be responsible for the biases found in the population level category emergence simulations. The

success of associative learning mechanisms in solving an increasing amount of speech acquisition tasks suggests that speech-specific cognitive category learning adaptations may not be necessary for speech learning.

2.2.1 Statistical learning of speech sound categories

Learnability is one constraint for the characteristics of speech sound categories, i.e. categories that are easier to learn and reproduce and will tend to be maintained, meaning general learning mechanisms shape language patterns (see e.g. Saffran, 2003; Oudeyer, 2005). However, speech learning is shaped by many more sources of information than heard acoustic speech. Computational models have shown that automatic discovery of the intuitive speech sound categories from speech only is very difficult, but using additional sources of information helps in discovering linguistically motivated speech sound categories, meaning multiple modalities have played a role in the discovery of the building blocks of speech in development.

If speech sound categories are searched from speech utterances alone, i.e. in an unsupervised manner where no information from other modalities is used to indicate phoneme identities or boundaries, computational algorithms fail to reach human-like performance in phonetic segmentation or annotation. Unsupervised segmentation of speech into phonemes based on statistical properties of speech spectra reaches about 70-85% accuracy (see Räsänen, 2012; Scharenborg, Wan, & Ernestus, 2010 for reviews). Labelling of correct phoneme categories is distorted by a lot of variation in phones depending on, for example, the surrounding phones and speaker identity. Many computational models of speech sound category acquisition (such as those of de Boer and Kuhl, 2003; Vallabha, McClelland, Pons, Werker and Amano, 2007) use heavily simplified learning data when compared to the natural continuous speech that language learning infants are exposed to. The functioning of these models cannot then be reliably compared to human performance.

Unsupervised methods for learning acoustic sub-word units from continuous speech (Varadarajan, Khundapur, and Dupoux, 2008; Lee and Glass,2012) successfully discover some speech patterns. However, the discovered patterns do not directly match with linguistically motivated phonemes, but are context sensitive language units such as allophones. As soon as correct category information is induced into the models (i.e. *supervised learning*), such as when phones are accompanied with their linguistically motivated labels while training, the learning accuracy comes close to the supervision criteria (e.g. Keshet, Shalev-Shwartz, Singer, & Chazan, 2005). It seems clear then, that acoustic regularities can be computationally discovered from speech signals, but the patterns found only classify nicely to linguistic categories (such as phonemes) if the categories are defined manually. Further, if information about the lexicon is included, then computational models start showing acquisition of phonological categories (e.g. Feldman, Myers, White, Griffiths, & Morgan, 2013; Thiessen, 2011; Swingley, 2009; ter Schure, Junge, & Boersma, 2016).

Using simplified vowel learning tasks and a Bayesian model, Feldman, Griffiths and Morgan (2009) showed that joint learning from lexical and distributional information of speech sounds leads to discovery of correct phonetic categories, whereas using distributional information alone tends to cluster overlapping phonetic categories together more easily. Having access to the referential meanings of words in continuous utterances may also facilitate the learning of word segmentation (Räsänen & Rasilo, 2015, see also François, Cunillera, Garcia, Laine, & Rodriguez-Fornells, 2016). Further, Martin, Peperkamp and Dupoux (2013) propose, using a computational simulation, that because of the large amount of allophonic variation in real speech, distributional information in the speech data alone makes phoneme discovery extremely difficult, and using word form knowledge facilitates solving the problem.

In addition to computational evidence, behavioral studies have shown that lexical (Thiessen, 200; White, Griffiths and Morgan, 2013), visual cues (Yeung & Werker, 2009; Teinonen, Aslin, Alku Csibra, 2008) and speech articulation (Majorano, Vihman & DePaolis,

2014) affect phonetic discrimination in human learners. Phonemes have also been argued to not be the basic units of perception, but rather context sensitive allophonic units (Mitterer, Scharenborg & McQueen, 2013; Reinisch, Wozny, Mitterer & Holt, 2014). Phonological awareness is also affected by acquiring literacy (see Anthony & Francis, 2005, for a review). It therefore seems that we are able to categorise speech sounds only because we have learnt the (context sensitive) category boundaries in speech, using several other sources of information that cue these otherwise seemingly arbitrary boundaries. When studying the emergence of speech sound categories in evolution, we should therefore not limit ourselves by focusing on intuitive speech representations, or studying characteristics of speech signals in isolation. Rather, we should look at the speech phenomena as a complete multimodal process.

2.2.2. Effects of speech production of category acquisition

In Section 2.2.1 we saw that learning to perceive speech categories probably depends on information from other modalities than speech itself. Since there would be no speech acquisition without acquiring the speech production skill as well, the learning of speech articulation must also impose biases to the structure of a language. Behavioural studies have shown that infants' babbles are shaped by non-vocalic and vocalic feedback by their caregivers, and caregivers are likely to reinforce more adult-like babbles (Goldstein, King, & West, 2003; Goldstein & Schwade, 2008). Also, infants' ability to imitate speech does not seem to be innate, but slowly develops during the first and second year, and importantly, seems to be tied to the *parents imitating their children* rather than the other way around (Gros-Louis et al., 2006; Kokkinaki & Kugiumutzakis, 2000; Jones, 2009; Kokkinaki & Vitalaki, 2013; Pawlby, 1977; Masur & Rodemaker, 1999). It therefore seems evident that infants do not learn a fully

functional language, including the speech production aspect, without interacting with other people (see also Sachs, Bard, & Johnson, 1981 for a case study).

Speech production is also complicated to learn because of infants' limited control of their articulatory systems. Children's articulation shows more variability compared to adults (Smith & Zelaznik, 2004; Walsh, Smith, & Weber-Fox, 2006), and reaches adult like motor control only at the age of 14 years (Smith & Zelaznik, 2004). Several computational models have modelled infant speech acquisition in interaction with caregivers in order to investigate the learning mechanisms behind this complex process. These methods mainly rely on associative learning where the infant associates its babbles into vocal responses by its caregiver (Miura, Katsushi, Yuichiro, & Minoru, 2008; Miura, Yoshikawa, & Asada, 2007; Howard & Messum, 2011; see the introduction of Rasilo and Räsänen, 2017, for a review of related studies). In the latest study by Rasilo & Räsänen (2017), a virtual infant learns to imitate the eight Finnish vowel sounds by learning to associating its (inaccurate) babbles (with an infant-sized vocal tract) with responses from adult human participants.

The details of the modeling processes show several interesting connections to the principles of speech sound category characteristics discussed in section 2.1. above. Due to the developing motor control and variation in infant articulation, learnability may increase when articulatory categories are further apart, so the category boundary is less likely to be accidentally crossed (c.f. *sensory discriminability* above), and when acoustic categories are maximally separated (c.f. *maximal perceptual contrast* above). Easy articulatory configurations may also be discovered faster in the vocal exploration (babbling) process (c.f. *minimal articulatory effort* above).

3. The Emergence of Combinatorial Structure

Above, we have discussed the emergence of speech categories, primarily in contexts where categories are treated as being in competition with each other in a signal space. However,

phonemes are very rarely used in isolation. In this section, we discuss the emergence of speech sounds as part of larger structures. Work on how structure emerged at the combinatorial level (meaningless building blocks combining) can inform us about how units of sound can be discretised from one another and reused in larger meaningful signals.

The emergence of structure on the combinatorial level has been subject to many hypotheses. Work on pressures for signals to be more learnable is reviewed in section 3.1. Work on pressures created by production constraints is summarised in section 3.2. Combinatorial structure as a solution to the problem of noise is discussed in section 3.3. Iconicity is discussed in section 3.4, and we the have a brief section on measuring combinatoriality in signals in section 3.5, before summarising in section 3.6.

Most the of the experiments discussed in this section are again artificial language learning or creation experiments which use continuous signal spaces from which discrete building blocks can emerge. These experiments use signal spaces that are distinct from actual speech apparatus, or indeed any other linguistic modality, in order to prevent interference from participants' existing linguistic knowledge.

3.1. Effects of Learnability in Transmission

As discussed in section 2.1, pressures for language to be learnable may greatly influence its structure. In this section, we will review studies which explore how learnability has influenced the emergence of combinatorial structure.

Del Giudice (2012) explored how the learnability of signals influences the structure that emerges in those signals. The experiment used a paradigm developed by Galantucci (2005); a continuous signal space designed to prevent the use of established symbols. The paradigm consists of a magnetic stylus on a graphics tablet, which participants use to create signals. The signal space moves vertically in a constant drift so that participants can only manipulate signals on the horizontal dimension. Del Giudice (2012) used the stylus in an iterated learning

experiment with transmission chains of between 6 and 9 generations. In this experiment there was no communication, but only learning and recall. Participants learnt a set of signal-meaning pairs and were then asked to recall the signals from memory. The study found that signals were much simpler and more learnable in later generations as a result of the transmission process. Del Giudice (2012) hypothesise that combinatorial structure is at the root of signals becoming more learnable, but struggled to make a solid claim because of the subjective nature of their analysis, using qualitative measures for structure rather than a more robust quantitative measure.

Verhoef (2012) also used an iterated learning experiment to explore the emergence of combinatorial structure. Participants used a slide whistle (or swanny whistle). A slide whistle has a piston so that the pitch of signals can be manipulated in a continuous fashion. Participants learnt a set of signals that had no meanings attached, and were then asked to recall the signals. Their output was then the next participant's input and so on for 10 generations. In this experiment, the first generation received a minimally structured repertoire to reproduce. Similar to the study of Del Giudice (2012), Verhoef (2012) found that signals became more structured, and that these more structured signal repertoires were more learnable than the initial repertoires. Signals tended to converge to using combinations of a small number of building blocks as the transmission chain progresses. These building blocks were separated by pauses in the signals and were things such a little peeps or larger whistle trajectories (e.g. from high to low pitch). More information about how this structure was measured is in section 3.5. Learnability was also found to be improve by showing a significant cumulative decrease in the recall error, defined as the distance between a learned signal and its reproduction.

3.2. Effects of Physical Constraints

While we found in the previous section that combinatorial can emerge with pressures for learning alone, it is still possible that other pressures may contribute to the speed or nature of its emergence. In section 2.2.2 we discussed how constraints of speech production might affect the

learning of speech categories. Here, we discuss how production constraints can affect the emergence of combinatorial structure in different ways.

Galantucci, Kroos and Rhodes (2010) used the stylus paradigm from Galantucci (2005) to investigate how *rapidity of fading* affects the structure of signals. Rapidity of fading is a signal's lack of permanence as it is produced. Speech has fast rapidity of fading as the signal disappears almost as soon as it is produced. In one condition in the experiment, signals disappeared as soon as the participant had finished drawing them (the fast fading condition). In the other condition, signals slowly disappeared off the screen for 2.5 seconds. In the fast fading condition, signals displayed more combinatorial reuse of signal elements.

Little and de Boer (2014) used the same slide whistle paradigm as in Verhoef (2012) in an iterated learning experiment, but to investigate the role of the size of signal space on combinatorial structure. Hockett's (1960) hypothesised that combinatorial structure emerged as the result of pressures for discrimination that occurs when an articulation space is saturated in the amount of distinctions it can make. Little and de Boer (2014) tested if a smaller signal space would make combinatorial structure emerge more quickly. They had one condition that was a replication of Verhoef (2012), but their other condition had a constrained signal space. They put a stopper on the piston of the slide whistle that prevented the participants from using the whole space, and found that with restricted signal space, there was a more consistent growth of combinatorial structure. Little, Eryılmaz, & de Boer (2016) have since looked at the effects of the dimensionality of a signal space, but this study is covered in section 3.5, as signal space dimensionality is tied up in the mappability of a signal space.

3.4. Effects of Noise

The amount of noise in a system is very related issue to the physical constraints above. That is, the size and dimensionality of a signal space affects how quickly a signal space gets crowded. Noise also affects how quickly a signal space gets crowded, which then has a knock

on effect on the emergence of combinatorial structure. Tria, Galantucci, & Loreto (2012) implemented a simulation to test the effect of noise on a set of agents trying to communicate a structured meaning space. They showed that the emergence of combinatoriality was dependent on the level of transmission noise, where higher levels of noise favoured more combinatorial strategies. The signals used were sequences of abstract, discrete items from an open set, so the results are perhaps not relevant to discussing the emergence of structure from an unstructured continuous space.

Zuidema and de Boer (2009), however, did use continuous signals to investigate the effect of noise. They focused on the relationship between the transmission bottlenecks caused by noisy communication, and combinatoriality. The limited transmission capability of the channel given the size of the system to be learned, whether it be caused by noise or some other constraint on transmission, is called a transmission bottleneck. Zuidema and de Boer (2009) show that the specific type and level of noise impacts the pressure for combinatoriality in a system that starts out random and holistic. The study represented signals as continuous trajectories in an abstract acoustic space. The repertoires of signals were optimised using two separate techniques, which produced comparable results. In one, a random initial repertoire was optimised, using a hill-climbing heuristic, for within repertoire distinctiveness. In the other, a population of agents optimised their random initial repertoires for communicative success in an imitation game. While the model produces qualitative (superficial) combinatorial structure under the right constraints (most notably with the presence of noise), the authors stress that they do not employ a quantitative measure of combinatoriality, citing the ambiguity of the terms building block and meaningful variation as some of the issues with developing one for the fine-grained trajectories they employ. We will discuss a little about the challenges of measuring combinatorial structure in section 3.5.

3.4 Effects of Iconicity and Conventionalisation

Another hypothesis for the emergence of combinatorial structure is that it exists to compensate for an inability to communicate information effectively in other ways, such as iconicity. Goldin-Meadow & McNeill (1999) argue that signs may need to lose their iconicity through conventionalisation before combinatorial structure can emerge through reanalysis of elements of a signal that are then perceived to be meaningless because of a lack of iconicity.

Roberts and Galantucci (2012) explicitly tested whether conventionalisation will promote combinatorial structure and argue that their results show that combinatorial structure comes from a process of conventionalisation. They used the stylus paradigm in a communication game where participants took it in turns to send and receive signals which referred to a set of animal shapes. They found that combinatorial structure negatively correlated with iconicity in the signals. Iconicity here was measured by having naive participants match signals to their intended meanings.

Roberts et al. (2015) also used the stylus paradigm to test directly whether iconicity impedes the emergence of combinatorial structure. They created different levels of available mappings between meanings and signals. In one condition, participants had to communicate meanings which were squiggly lines, which could be directly represented by the paradigm, or they were asked to communicate green circles which were very difficult to represent iconically. Indeed, they found that combinatorial structure was more prevalent in signals where there was no obvious available mapping between signal and meaning space.

Verhoef, Kirby, and de Boer (2015) tested a similar hypothesis using the slide whistle paradigm in an iterated learning experiment. Signals referred to a set of novel object meanings. In one condition, the output signal-meaning pairs were taught to the next participant with no manipulation. In the other condition, the output signals from the person before were used as the input, but the meanings they were paired with were randomly assigned. This second condition was designed to minimise the amount of iconicity in the signals, as the meaning the signal referred to changed after every generation. They found that the emergence of combinatorial

structure was delayed by having consistent signal-meaning pairs because of a reliance on iconic strategies.

Little, Eryılmaz, & de Boer (2015, 2016) have since explored the relationship between iconicity and combinatorial structure using the Leap Motion paradigm (Eryılmaz and Little, 2016). This paradigm utilises a Leap Motion sensor, which is an infrared motion detector that maps hand movements in the space above it and transforms hand-positions into audio output. This audio output can be manipulated on any continuous feature (e.g. pitch, volume) by moving a hand on the vertical dimension or horizontal dimension. In most experiments only pitch is manipulated on the horizontal dimension, but some experiments (e.g. Little et al., 2015) used both pitch and volume attached to different spatial dimensions.

Little, Eryılmaz and de Boer (2016) explored more nuanced grades of available mappings than previous studies, in the form of dimensionality mismatches. Signals only had one continuous dimension (pitch), but participants had to describe squares from continuous meaning spaces with increasing dimensionality. Initially, squares only differed in size, which can be mapped to pitch with relative ease (e.g. low sounds referring to bigger squares). The meaning space then expanded to squares that differed in size and shade of orange (controlled by the physical distance on the RGB ratio of green to red) and finally they differed in size, shade and darkness. The idea was that as the meaning space increased in its dimensionality, this would make transparent mappings with the one dimensional signal space less and less feasible. The results showed that signals from repertoires used to label squares that differ along more dimensions had hallmarks of combinatorial structure (an increase in signal duration and the amount of movement in signals). This study was born from predictions made by a computational model (de Boer & Verhoef, 2012) which also found that combinatorial strategies were needed when the meaning space had more dimensions than the signal space.

Little et al. (2015) further explored different ways in which iconicity could be disrupted. Again using the Leap Motion paradigm, in one condition, participants created continuous

signals for meanings which differed along continuous dimensions (e.g. size), and in the other condition participants created continuous signals for meanings which differed along discrete dimensions (e.g. texture). They found that there was significantly more movement in signals made for the meanings in the discrete condition than in the continuous condition, which displayed no signs of combinatorial structure.

Importantly, this section focuses on a phenomenon that is hypothesised to inhibit, rather than facilitate, the emergence of combinatorial structure. There are some languages in the world which exist without a level of combinatorial structure, including Al-Sayyid Bedouin Sign Language (Sandler et al., 2011) and Central Taurus Sign Language (Caselli et al., 2014), which have no minimal pairs. Interestingly, these languages are emerging sign languages. Under the well-established assumption that sign languages have more iconicity available to them than spoken languages, then this availability for iconicity may explain the absence of iconicity in these languages, but perhaps not explain its presence in spoken languages.

3.5. Measuring structure

One consistent problem with all of the studies we have discussed here that use continuous signal spaces, is that of measuring the structure that comes out of these studies. Here, we discuss some of the measures used to measure structure in signals produced by some of the paradigms mentioned in section 3.

With the slide whistle paradigm, analysis relies on both qualitative and quantitative measures for combinatorial structure. One measure used by Verhoef, Kirby and de Boer (2014), which is highly similar to that used by Verhoef (2012), is to segment the whistle signals, and have the segmented forms grouped by hierarchical clustering to detect variations of the same prototypical form. More specifically, they segmented the signals using pauses in the signals as delimiters. The similarity, or rather the dissimilarity, between signals was measured using Derivative Dynamic Time Warping algorithm on each pair of signals. After segmentation, they

used average linkage hierarchical agglomerative clustering to find segments that are so similar that they can be considered instances of the same prototype, until no two clusters remained that were more similar than a certain threshold. The number of occurrences for the prototype segments were then used to calculate the entropy of the repertoire (Shannon, 2001). Since entropy is a measure of compressibility, and since reusing components in a signal increases its compressibility, one can use entropy as an index of structure. Verhoef, Kirby and de Boer (2014) also calculate "associative chunk strength", which is an index of how sensitive the building blocks are to ordering, calculated by first averaging the bigram and trigram frequencies over the input set. Then the associative chunk strength of an utterance can be calculated by summing the strengths of its constituent chunks.

Another measure for combinatorial structure has been developed for several experiments using the stylus paradigm (Galantucci et al., 2010; Roberts & Galantucci, 2012; Roberts et al., 2015). Within these studies, as with the slide whistle signals, while the signals are continuous, the measure for combinatoriality depends on form boundaries. The signals were segmented into their parts using intentional gaps (left by lifting the stylus). The signals were represented by the mean stylus position within a signal, the number of crossings the stylus made of that mean point in them, and the proportions of the portions between these crossings.

In Little et al. (2015), instead of modeling individual trajectories, the authors opt for modeling the repertoire of trajectories. In order to try and model combinatorial structure, the repertoires are used as the training data for a range of Hidden Markov Models (HMMs) with different numbers of states. The number of states directly corresponds to the number of discrete building blocks, providing a data-driven way to discretise the data, including temporal dependencies. The model with the lowest Bayesian Information Criterion score is chosen as the one with an ideal number of states. In the study of Little et al. (in review), meaning spaces with more dimensions were shown to require signals with more states to achieve the same level of

performance as signals for meaning spaces with fewer dimensions, showing a link between participant behaviour and this parameter, supporting the validity of the model.

3.6. Summary

The experimental and modeling work on the emergence of combinatorial structure reviewed here shows that there are multiple pressures for the emergence of combinatorial structure. These include the relative robustness of combinatorial structure against certain kinds of noise and signal space crowding, and the learnability of combinatorial communication systems. While we have not explicitly discussed the pressure for expressivity in this section, this is something that ties directly in with signal space crowding, in that crowding creates a pressure for signals to find ways to be distinct in the say way that expressivity does.

5. Summary and Conclusions

In this review, we have focused on experimental and computational studies of emergence of speech structure, involving mechanisms underlying the emergence and discovery of meaningless "building blocks" of speech (such as phonemes or allophones) and mechanisms underlying their combination of those building blocks into larger speech structures (such as syllables or words).

Summarising from the sections above, in population level simulations, articulatory effort, auditory and articulatory sensory discriminability, and lexical and language transmission constraints have been shown to play a part in emergence of human language-like phonetic systems. When it comes to *discovery* of speech sound categories from spoken speech utterances, thus far, computational studies of unsupervised acquisition of phone categories or unsupervised segmentation of continuous speech into phones have failed to provide accuracies comparable to

human performance in the tasks. When supervision is included, computational models start to approach human performance. Discriminatory information (such as lexical constraints) has been shown to aid phonetic acquisition in computational as well as experimental studies. Additionally, interaction with caregivers/peers has a major influence in phonetic acquisition, indicating that category structure in speech is not learned from linguistic input alone, but emerges in complex interaction with the learner's environment.

Experimental studies looking at population-level processes on the the categorical emergence of speech categories is somewhat thin on the ground, though work looking at how both learning and communication can affect the emergence of semantic categories is potentially relevant. It seems that individuals may be able to design optimal categories alone, but communication becomes a problem when people's individual categorisation preferences do not match. It is then the learning and transmission of a category system, as well as the pressure for expressivity, that makes a category system optimal within a population. Though this has been shown with semantic-spaces, we feel that the same principles will play out in the context of the categorisation of speech sounds, as it is in line with some of the population level modelling work, e.g. Liljencrants & Lindblom (1972) showing the importance of an expressive system (maximally distinct vowel categories) and de Boer (2000) showing the importance of speech categories being learnable.

Regarding the combinatorial nature of language, experimental and computational studies have shown that pressures for discriminability, production constraints, learnability, noise levels and lack of permanence (or rapidity of fading), have been shown to lead to more combinatorial structure, while iconicity has been found to inhibit combinatorial structure. Interestingly, several connections can be drawn from these findings to the infant speech learning literature. For instance, the size of a signal space repertoire may cause pressures during development. The initial amount of important concepts that should be discriminable for a small infant is rather limited, consisting of people or objects that are necessary in simple everyday situations (e.g.

daddy, mommy, baabaa and bye might be the first words spoken by children, see Tardif et al., 2008). If infants are to discriminate reliably between acoustic realisations and articulatory productions of a small set of concepts, it would be beneficial that their sub-word units should be far away from each other both acoustically (to minimise errors while hearing words) and in articulatory sense (to minimise errors while producing words). When the amount of important concepts gradually grows, requiring more discriminatory signals, it may be a more efficient learning strategy to start combining the robust set of building blocks already learned, rather than unlearn and reorganise the auditory and articulatory adaptations for a larger number of building blocks. Also, as young infants do not have accurate control over their articulatory systems (Smith & Zelaznik, 2004), the restricted and noisy signal (or articulatory) space may encourage combining these simple building blocks of speech during the learning process.

As a concluding note, we see work done on the emergence of both categorical and combinatorial structure to be very much in its infancy and foresee this kind of work exploding as new experimental and computational methods are developed. In particular, there is a current gap in experimental work done on the categorical emergence of speech categories. Further, while we have identified many things which we can say "have an effect" on the emergence of combinatorial structure, much work needs to be done in order to untangle what pressures are more influential than others. Experimental and computational models with different conditions are an excellent tool to help us answer questions like this. For example, the question of whether learnability or expressivity is more influential has been addressed using these methods in recent work on the emergence of compositional structure (e.g. Carr er al., 2016; Kirby et al., 2015), but a gap remains in the literature regarding combinatorial structure.

6. Acknowledgements

Financial support from the ERC starting grant, ABACUS, project number 283435. HR is additionally supported by the Academy of Finland project titled "Computational Modeling of Language Acquisition", and the Finnish Cultural Foundation.

7. References

Blevins, J. (2004). *Evolutionary Phonology.* Cambridge University Press: Cambridge.

Canini, K. R., Griffiths, T. L., Vanpaemel, W., & Kalish, M. L. (2014). Revealing human inductive biases for category learning by simulating cultural transmission. *Psychonomic Bulletin & Review*, *21*(3), 785–793.

Carr, J. W., Smith, K., Cornish, H., & Kirby, S. (2016). The Cultural Evolution of Structured Languages in an Open‐Ended, Continuous World. *Cognitive science.* doi:10.1111/cogs.12371

Carré, R. (1996). Prediction of vowel systems using a deductive approach. In H. T. Bunnell & W. Idsardi (Eds.), *Proceedings of Fourth International Conference on Spoken Language Processing. ICSLP '96* (pp. 1593–1597). IEEE.

Caselli, N., Ergin, R., Jackendoff, R., & Cohen-Goldberg, A. (2014). The emergence of phonological structure in central taurus sign language. From *Sound to Gesture.* Padua, Italy.

de Boer, B. (2000). Emergence of Sound Systems Through Self-Organisation. In C. Knight, M. Studdert- Kennedy, & J. R. Hurford (Eds.), *The Evolutionary emergence of language: Social function and the origins of linguistic form* (pp. 177-198). New York : Cambridge University Press.

de Boer, B., & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online: ARLO*, *4*, 129-134.

de Boer, B., Sandler, W., & Kirby, S. (2012). New perspectives on duality of patterning: Introduction to the special issue. *Language and Cognition*, *4*(04), 251–259.

de Boer, B., & Verhoef, T. (2012). Language dynamics in structured form and meaning spaces. *Advances in Complex Systems. A Multidisciplinary Journal*, *15*(3), 1150021–1–1150021–20.

Del Giudice, A. (2012). The emergence of duality of patterning through iterated learning: Precursors to phonology in a visual lexicon. *Language and Cognition*, *4*(4), 381–418.

Eryılmaz, K., & Little, H. (2016). Using leap motion to investigate the emergence of structure in speech and language. *Behavioral Research Methods*. doi:10.3758/s13428–016–0818–x

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). Learning phonetic categories by learning a lexicon. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the cognitive science society (CogSci 2009)* (pp. 2208–2213).

Feldman, N. H., Myers, E. B., White, K. S., Griffiths, T. L., & Morgan, J. L. (2013). Word-level information influences phonetic learning in adults and infants. *Cognition*, *127*(3), 427–438.

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

François, C., Cunillera, T., Garcia, E., Laine, M., & Rodriguez-Fornells, A. (2016). Neurophysiological evidence for the interplay of speech segmentation and word-referent mapping during novel word learning. *Neuropsychologia*.

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science*, *29*(5), 737–767.

Galantucci, B., Kroos, C., & Rhodes, T. (2010). The effects of rapidity of fading on communication systems. *Interaction Studies*, *11*(1), 100–111.

Goldin-Meadow, S., & McNeill, D. (1999). The role of gesture and mimetic representation in making language the province of speech. In M. C. Corballis, & S. Lea (Eds.) *Evolution of the Hominid Mind,* (pp. 155–172). Oxford University Press.

Goldstein, M. H., King, A. P., & West, M. J. (2003). Social interaction shapes babbling: testing parallels between birdsong and speech. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(13), 8030–8035.

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science*, *19*(5), 515–523.

Gros-Louis, J., West, M. J., Goldstein, M. H., & King, A. P. (2006). Mothers provide differential feedback to infants' prelinguistic sounds. *International Journal of Behavioral Development, 30*(6), 509–516.

Hockett, C. F. (1960). The origin of speech. *Scientific American*, *203*, 88–111.

Howard, I. S., & Messum, P. (2011). Modeling the development of pronunciation in infant speech acquisition. *Motor Control*, *15*(1), 85–117.

Jones, S. S. (2009). The development of imitation in infancy. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*(1528), 2325–2335.

Keshet, J., Shalev-Shwartz, S., Singer, Y., & Chazan, D. (2005). Phoneme alignment based on discriminative learning. In *Proceedings of INTERSPEECH 2005* (pp. 2961–2964).

Kirby, J. & Sonderegger, M. (2015) Bias and population structure in the actuation of sound change. *arXiv:*1507.04420 [physics].

Kirby, S. (2002b). Natural language from artificial life. *Artificial Life,* 8(2):185-215.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition,* 141, 87-102.

Kokkinaki, T., & Kugiumutzakis, G. (2000). Basic aspects of vocal imitation in infant-parent interaction during the first 6 months. *Journal of Reproductive and Infant Psychology*, *18*(3), 173–187.

Kokkinaki, T., & Vitalaki, E. (2013). Comparing spontaneous imitation in grandmother-infant and mother-infant interaction: a three generation familial study. *International Journal of Aging & Human Development*, *77*(2), 77–105.

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *363*(1493), 979–1000.

Lee, C.-Y., & Glass, J. (2012). A nonparametric Bayesian approach to acoustic model discovery. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1 (pp. 40–49). Stroudsburg, PA: Association for Computational Linguistics

Liljencrants, J., & Lindblom, B. (1972). Numerical Simulation of Vowel Quality Systems: The Role of Perceptual Contrast. *Language*, *48*(4), 839-862.

Lindblom, B., Macneilage, P., & Studdert-Kennedy, M. (1984). Self-organizing processes and the explanation of phonological universals. In B. Butterworth, B. Comrie, & Ö. Dahl (Eds.), *Explanations for Language Universals* (pp. 181–204). DE GRUYTER.

Little, H., & de Boer, B. (2014). The effect of size of articulation space on the emergence of combinatorial structure. In A. Cartmill Erica, S. Roberts, H. Lyn, & H. Cornish (Eds.), *The Evolution of Language: Proceedings of the 10th international conference (EVOLANGX)* (pp. 479–481). World Scientific.

Little, H. & Eryılmaz, K. (2016) Using leap motion to investigate the emergence of structure in speech and language. *Behaviour Research Methods.* doi:10.3758/s13428-016-0818-x

Little, H., Eryılmaz, K., & de Boer, B. (In review). Signal dimensionality and the emergence of structure.

Little, H., Eryılmaz, K., & de Boer, B. (2015). Linguistic modality affects the creation of structure and iconicity in signals. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *The 37th annual meeting of the Cognitive Science Society (CogSci 2015)* (pp. 1392–1398). Austin, TX: Cognitive Science Society.

Little, H., Eryılmaz, K., & de Boer, B. (2016). Differing Signal-meaning Dimensionalities Facilitates The Emergence Of Structure. In S. G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Feher, & T. Verhoef (Eds.), *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANGX1)* (pp. 182–190).

Majorano, M., Vihman, M. M., & DePaolis, R. A. (2014). The relationship between infants' production experience and their processing of speech. *Language Learning and Development*, *10*(2), 179-204.

Martin, A., Peperkamp, S., & Dupoux, E. (2013). Learning phonemes with a proto-lexicon. *Cognitive Science*, *37*(1), 103–124.

Masur, E., & Rodemaker, J. E. (1999). Mothers' and infants' spontaneous vocal, verbal, and action imitation during the second year. *Merrill-Palmer Quarterly, 45*(3), 392–412.

Matthews, C. (2009). *The emergence of categorization: Language transmission in an Iterated Learning Model using a continuous meaning space* (MSc Thesis). University of Edinburgh .

McDonough, J., Ladefoged, P., & George, H. (1992). Navajo vowels and universal phonetic tendencies. *The Journal of the Acoustical Society of America*, *92*(4), 2416.

Mitterer, H., Scharenborg, O., & McQueen, J. M. (2013). Phonological abstraction without phonemes in speech perception. *Cognition*, *129*(2), 356-361.

Miura, K., Katsushi, M., Yuichiro, Y., & Minoru, A. (2008). Realizing being imitated: Vowel mapping with clearer articulation. In *2008 7th IEEE International Conference on Development and Learning*. (pp. 262-267), http://doi.org/10.1109/devlrn.2008.4640840

Miura, K., Yoshikawa, Y., & Asada, M. (2007). Unconscious anchoring in maternal imitation that helps find the correspondence of a caregiver's vowel categories. *Advanced Robotics: The International Journal of the Robotics Society of Japan*, *21*(13), 1583–1600.

Oudeyer, P. Y. (2005). How phonological structures can be culturally selected for learnability. *Adaptive Behavior*, *13*(4), 269-280.

Pawlby, S. J. (1977). Imitative interaction. In H. Schaffer (Ed.), *Studies in mother-infant interaction* (pp. 203–224). New York: Academic Press.

Perfors, A., & Navarro, D. J. (2014). Language evolution can be shaped by the structure of the world. *Cognitive Science*, *38*(4), 775–793.

Rasilo, H., & Räsänen, O. (2017). An online model for vowel imitation learning. *Speech Communication*, *86*, 1-23.

Reinisch, E., Wozny, D. R., Mitterer, H., & Holt, L. L. (2014). Phonetic category recalibration: What are the categories? *Journal of phonetics, 45,* 91-105.

Räsänen, O. (2012). Computational modeling of phonetic and lexical learning in early language acquisition: Existing models and future directions. *Speech Communication*, *54*(9), 975–997.

Räsänen, O. & Rasilo, H. (2015). A joint model of word segmentation and meaning acquisition through cross-situational learning. *Psychological Review*, *122*(4), 792–829.

Roberts, G., Galantucci, B. (2012). The emergence of duality of patterning: Insights from the laboratory. *Language and Cognition 4(*4), 297-318.

Roberts, G., Lewandowski, J., & Galantucci, B. (2015). How communication changes when we cannot mime the world: Experimental evidence for the effect of iconicity on combinatoriality. *Cognition*, *141*, 52–66.

Sachs, J., Bard, B., & Johnson, M. L. (1981). Language learning with restricted input: Case studies of two hearing children of deaf parents. *Applied Psycholinguistics*, *2*(01), 33-54.

Saffran, J. R. (2003). Statistical language learning mechanisms and constraints. *Current directions in psychological science*, *12*(4), 110-114.

Sandler, W., Aronoff, M., Meir, I., & Padden, C. (2011). The gradual emergence of phonological form in a new language. *Natural language & linguistic theory,* 29(2), 503–543.

Scharenborg, O., Wan, V., & Ernestus, M. (2010). Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries. *The Journal of the Acoustical Society of America*, *127*(2), 1084-1095.

Shannon, C. E. (2001). A Mathematical Theory of Communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, *5*(1), 3–55.

Silvey, C., Kirby, S., & Smith, K. (2013). Communication leads to the emergence of sub-optimal category structures. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.) *Proceedings of the 35th annual conference of the Cognitive Science Society* (pp. 1312–1317). Austin, TX: Cognitive Science Society.

Silvey, C., Flaherty, M., Goldin-Meadow, S., Kirby, S., & Smith, K. (2016). Communication without a language model inhibits the emergence of systematic structure. In S. G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Feher, & T. Verhoef (Eds.), *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANGXI)*. doi:10.17617/2.2248195.

Smith, A., & Zelaznik, H. N. (2004). Development of functional synergies for speech motor coordination in childhood and adolescence. *Developmental Psychobiology*, *45*(1), 22–33.

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*(6640), 381–382.

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*(1536), 3617–3632.

Tardif, T., Fletcher, P., Liang, W., Zhang, Z., Kaciroti, N, & Marchman, V. A. (2008). Baby's first 10 words. *Developmental Psychology*, *44*(4), 929–938.

Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition*, *108*(3), 850-855.

ten Bosch, L. (1995). Lexically-Based vowel dispersion: a case study for Dutch. *Proceedings of the Institute of Phonetic Sciences, University of Amsterdam, 19,* pp. 39-50.

ter Schure, S. M. M., Junge, C. M. M., & Boersma, P. P. G. (2016). Semantics guide infants' vowel learning: Computational and experimental evidence. *Infant Behavior & Development*, *43*, 44–57.

Thiessen, E. D. (2007). The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language*, *56*(1), 16-34.

Thiessen, E. D. (2011). When variability matters more than meaning: the effect of lexical forms on use of phonemic contrasts. *Developmental Psychology*, *47*(5), 1448–1458.

Tria, F., Galantucci, B., & Loreto, V. (2012). Naming a structured world: a cultural route to duality of patterning. *PloS One*, *7*(6), e37744.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, *104*(33), 13273–13278.

van der Ham, S., & de Boer, B. (2015). Cognitive Bias for Learning Speech Sounds From a Continuous Signal Space Seems Nonlinguistic. *I-Perception*, *6*(5), 2041669515593019.

Varadarajan, B., Khundapur, S., & Dupoux, E. (2008). Unsupervised learning of acoustic sub-word units. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies Short Papers - HLT '08* (pp. 165–168). Ohio, USA.

Vaux, B., & Samuels, B. (2015). Explaining vowel systems: dispersion theory vs natural selection. *The Linguistic Review*, *32*(3). http://doi.org/10.1515/tlr-2014-0028

Verhoef, T. (2012). The origins of duality of patterning in artificial whistled languages. *Language and Cognition*, *4*(04), 357–380.

Verhoef, T., Kirby, S., & de Boer, B. (2014). Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals. *Journal of Phonetics*, *43*(1), 57–68.

Verhoef, T., Kirby, S., & de Boer, B. (2015). Iconicity and the Emergence of Combinatorial Structure in Language. *Cognitive Science*. http://doi.org/10.1111/cogs.12326

Walsh, B., Smith, A., & Weber-Fox, C. (2006). Short-term plasticity in children's speech motor systems. *Developmental Psychobiology*, *48*(8), 660–674.

Wedel, A., Kaplan, A., & Jackson, S. (2013). High functional load inhibits phonological contrast loss: a corpus study. *Cognition*, *128*(2), 179–186.

Werker, J. F., & Curtin, S. (2005). PRIMIR: A Developmental Framework of Infant Speech Processing. *Language Learning and Development: The Official Journal of the Society for Language Development*, *1*(2), 197–234.

Yeung, H. H., & Werker, J. F. (2009). Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. *Cognition*, *113*(2), 234-243.

Zuidema, W., & Boer, B. D. (2009). The evolution of combinatorial phonology. *Journal of Phonetics*, 37(2), 125-144