Dispersion, communication, and alignment: An experimental study of the emergence of structure in combinatorial phonology

Gareth Roberts*[1] and Robin Clark[1]

[1]Department of Linguistics, University of Pennsylvania

Abstract

Languages exhibit structure at a number of levels, including at the level of phonology, the system of meaningless combinatorial units from which words are constructed. Phonological systems typically exhibit greater dispersion than would be expected by chance. Several theoretical models have been proposed to account for this, and a common theme is that such organization emerges as a result of the competing forces acting on production and perception. Fundamentally, this implies a cultural evolutionary explanation, by which emergent organization is an adaptive response to the pressures of communicative interaction. This process is hard to investigate empirically using natural-language data. We therefore designed an experimental task in which pairs of participants play a communicative game using a novel medium in which varying the position of one's finger on a trackpad produced different colors. This task allowed us to manipulate the alignment of pressures acting on production and perception. Here we used it to investigate (a) whether above-chance levels of dispersion would emerge in the resulting systems, (b) whether dispersion would correlate with communicative success, and (c) how systems would differ if the pressures acting on perception were misaligned with pressures acting on production (and which would take precedence). We found that above-chance levels of dispersion emerged when pressures were aligned, but that the primary driver of communicative success was the alignment of production and perception pressures rather than dispersion itself. When they were misaligned, participants both found the task harder and (driven by perceptual demands) created systems with lower levels of dispersion.

Keywords: cultural evolution; phonology; combinatoriality; emergence of structure; language; communication; experiment

*Correspondence should be sent to Gareth Roberts, Department of Linguistics, University of Pennsylvania, 3401-C Walnut Street, Suite 300, Philadelphia, PA, 19104, USA. E-mail: gareth.roberts@ling.upenn.edu

In this paper we present an experimental investigation of how organizational structure, in particular dispersion, emerges in phonological systems. We focus on the role of communicative interaction and the alignment of production demands with perceptual demands as a cultural evolutionary mechanism for the emergence of dispersion.

1 Structure in phonology

A phonology is a set of discrete meaningless linguistic units (*phonemes*) that are recombined into meaningful units (*morphemes*), giving languages a two-layer structure known as *duality of patterning* (Hockett, 1960; Ladd, 2012). In English, for instance, the phonemes /a, s, k, n/ can be recombined into such words as *ask, can, knack*, and *snack*. Any given phoneme can be considered to correspond to representations in two distinct spaces. The first of these is the production space, defined in terms of which articulators are involved and how they interact. A vowel or consonant from a spoken language, for instance, can be defined in terms of such variables as the position of the tongue relative to the top or bottom of the mouth, the shape of the lips, or the degree and timing of vocal-fold constriction (Ladefoged & Johnson, 2015). Phonological units in sign languages can, similarly, be defined in terms of such variables as handshape, hand location, or direction of movement (Brentari, 2011). The second space is the perceptual space, defined in terms of the phoneme's perceptual form. For vowels and consonants, this perceptual form is acoustic, and a phoneme might be defined in terms of the components of its waveform. In sign languages there is a less clear-cut distinction between the articulatory and perceptual properties of phonemes, since viewers perceive such features as handshape and relative location rather directly. Nonetheless, there are still important differences between the spaces – for instance, differences of perspective mean that the signer's visual experience of their own signs is not the same as that of their interlocutors. It is also important to note that, while there is a clear causal relationship between the production space and the perceptual space, this does not mean that the relationship between the two spaces is intuitive or straightforward. It is well known to linguistics instructors that being able to reliably reproduce a particular speech sound does not entail even a basic conscious understanding of how it is produced.

While phonologies consist of discrete units, both the production and perceptual spaces are continuous. It has nonetheless long been noted that natural language phonologies are systematic, or structured, in the sense that phonemes do not appear to have been sampled at random from the continuous spaces. Rather, they seem to be well dispersed and symmetrical (de Boer, 2000). The purpose of this paper is to present a novel experimental paradigm for investigating where this organization comes from, as a result of interaction between the pressures acting on production on the one hand and those acting on perception on the other. The nature of the organization, as well as the role of perception and production, can be particularly clearly demonstrated in the case of vowels.

1.1 Vowel spaces

A vowel is a type of speech sound produced by allowing air from the lungs to pass through the vocal tract unobstructed. Vowel quality is varied by varying the shape of the vocal tract, principally by moving the tongue and lips. This leads to variation in formant frequencies, prominent concentrations of acoustic energy directly related to vocal tract resonances. Several formants matter for speech perception, but the first and second have traditionally been seen as the most important.* Standardly, vowel phonemes are plotted in a two-dimensional space with the y axis corresponding to the first formant (F_1, shown as increasing in frequency from top to bottom), and the x axis to the second formant (F_2, shown as increasing from right to left; see Fig. 1a). These values correspond well enough to tongue position that vowels in the top left of the space are standardly referred to as high front vowels, as they are produced by raising the tongue towards the top front of the mouth (for a more detailed introduction to the phonetics of vowels see Ladefoged & Johnson, 2015).

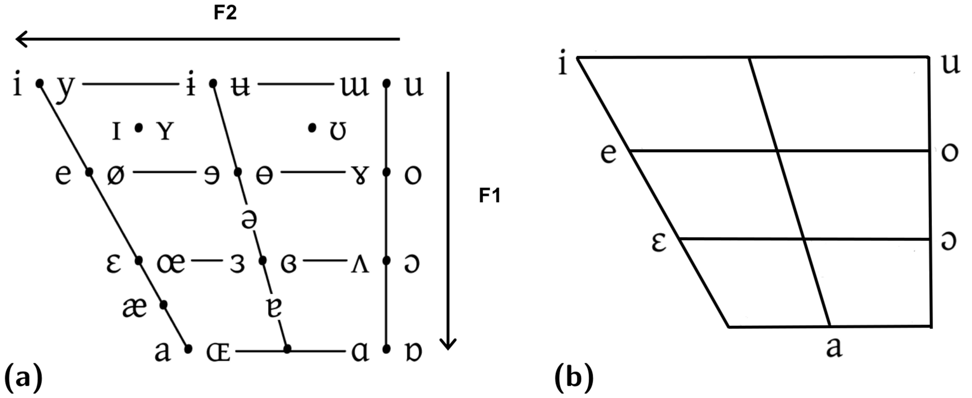

Figure 1: (a) Chart of the vowels of the world's languages (based on charts produced by the International Phonetic Association), with first and second formants indicated. The chart corresponds roughly to the mouth of a speaker facing left. For each pair of vowels, the vowel on the right is produced with lip rounding. (b) The vowels of Italian, a classic seven-vowel system.

Vowel charts for particular languages locate individual vowels in specific fixed locations (Fig. 1b shows the vowels of Italian). This is an idealization, driven in part by the realization that phoneme inventories can typically be described in terms of small number of parameterized features (Chomsky & Halle, 1968). First, and perhaps most obviously, it is an idealization across multiple speakers whose pronunciations vary with respect to each other. However, even for a single speaker, any phoneme corresponds to a cluster of phonetic realizations that vary both between the words they occur in and different instances of producing the same word. This has more than one consequence. First, it renders non-trivial any attempt to identify what precise acoustic or articulatory coordinates to attribute to a given phoneme, or even to identify whether two sounds correspond to different phonemes.

*Other formants can play a crucial role in speech perception, however. de Boer (2000), for example, captured this by calculating F_2 as the weighted average of higher formants (see also Mantakas, Schwartz, & Escudier, 1986).

Second, it means phonologies are dynamic evolving systems, whose units shift locations over time relative to one another (see, e.g., Gordon, 2002).

However, saying that a system is dynamic and evolving is different from saying it is not ordered. In fact, it has long been recognized that vowel systems tend to be efficiently dispersed and symmetrical (de Boer, 2000; Liljencrants & Lindblom, 1972; Schwartz, Boë, Vallée, & Abry, 1997), and the purpose of this study is to shed light on the cultural-evolutionary forces behind the order. In particular, how do interactive processes, in which individuals produce and perceive signals for communicative purpose, contribute to the dispersion of phonemes across the available space? We investigated this using an experimental study in which pairs of participants played a communication game using a novel communication medium. While our study was largely exploratory, we also manipulated the extent to which the communicative pressures acting on production and those acting on perception were aligned.

Our research questions and experimental approach will be laid out in Sections 3 and 4, and our method will be described in detail in Section 5. In Section 2 we will discuss theoretical models of vowel space organization.

2 Explaining phonological organization

Several different categories of theory have been proposed to account for the observation that vowel spaces exhibit structural organization (for reviews see de Boer, 2001; Vaux & Samuels, 2015). One category of theory starts from the observation that phoneme inventories can be described in terms of a small set of parameterized "distinctive features". For instance, the distinction between Italian /i, e, ɛ/ and /u, o, ɔ/ (Fig 1b) can be described in terms of the binary feature [+/- front], such that the former vowels are all [+front] and the latter are all [-front] (Chomsky & Halle, 1968). Classical markedness-based accounts of vowel spaces treat such distinctive features as having cognitive reality and an explanatory role (Chomsky & Halle, 1968; Jakobson & Halle, 1956). Some features are claimed to be more "marked" than others, and optimizing a vowel system means reducing its markedness and reducing the number of features needed to describe it. What precisely is meant by marked is less easy to answer than might be thought (de Boer, 2001, pp. 10–11; Haspelmath, 2006), and early work was characterized by relatively little attention to the details of markedness (Blevins, 2004, pp. 74–78), but the general idea is that marked features are typologically less common, more complex, or further from the default in some sense. An important criticism is that markedness is really a descriptive rather than explanatory account that thus risks circularity (Blevins, 2004; de Boer, 2001). That is, it is not obvious how markednesss can be derived from the physical and cognitive details of speech production and perception. Despite this concern, several researchers have developed approaches that attempt to identify a physical, real-world basis for these distinctive features (e.g., Flemming, 2001; Kager, 1999, pp. 10–11).

Two such approaches are *quantal theory* (Stevens, 1989; Stevens & Keyser, 2010) and the *distinctive region model* (Carré, Divenyi, & Mrayati, 2017; Mrayati, Carré, & Guérin, 1989), which ground distinctive features and markedness in the observation that the articulatory space is unevenly structured, such that small articulatory movements have greater acoustic effects in some regions of the space than others. Optimization in these models involves reducing articulatory effort for the speaker, while maintaining perceptual distinctiveness.

The accounts described so far can all be understood as being concerned with constraints on individual vowels (de Boer, 2001; Lindblom & Engstrand, 1989). Dispersion theory differs by focusing on the relationship *between* vowels (Lindblom, 2003). An optimal system is then one in which vowels are maximally dispersed. This model too makes reference to competition between articulation and perception, but with implicit emphasis on perception. Reducing production effort is not inherently consistent with maintaining perceptual distinctiveness (indeed, production effort could be maximally reduced by reducing all vowels to one mid-central vowel), so maximizing distinctiveness is most obviously driven by perceptual pressures. However, *maximal* dispersion to the corners and outer edges of the vowel space is not in fact necessary from a perceptual point of view; the vowels in only somewhat dispersed systems may still be sufficiently distinct for communication purposes. Part of the explanation for dispersion may thus be that the peripheries of the space are easier to locate reliably and are so likely to be preferred for articulatory as well as perceptual reasons.[†]

All the models we have discussed have something clear to say about what the design constraints are on phonological systems, whether these constraints come from inherent features of the vocal-auditory system or the language faculty, or from interaction between units in the system. However, phonological systems in natural language are not *designed* in the everyday sense of the term. Nor do they simply self-organize independently of human action. Rather, organized systems evolve as a result of a population of humans attempting to use them to achieve their ends (cf. Keller, 2005). In particular, people use combinations of phonemes to communicate meanings to each other. It is this interactive process that imposes the articulatory and perceptual constraints on the system. It thus seems reasonable to suppose that (along with transmission to learners; Kirby, Griffiths, & Smith, 2014) communicative interaction is a key mechanism by which phonological units settle into a conventional and organized system. The goal of this paper is to investigate this cultural evolutionary process experimentally and, by manipulating the extent to which they are aligned, distinguish the roles of perception and production as drivers of organizational structure.

3 Goals and research question

The fundamental goal of our study is to establish a new experimental approach for investigating phonological organization. However, the particular experiment presented here was designed to do three things. The first was to test whether a novel visual communication system would, through cooperative communication, begin to exhibit organization (in particular, dispersion and symmetry) at a greater than chance level. The second was to test whether dispersion correlates with communicative success. The third was to manipulate the

[†] Perceptual distinctiveness alone also fails to account for the fact that equivalently well dispersed systems are not equally common (Schwartz, Boë, Vallée, & Abry, 1997). One solution to this was the development of dispersion-focalization theory, combining dispersion theory with quantal theory (Schwartz, Boë, et al., 1997) to better take into account the nonlinear nature of the articulatory and perceptual systems.

extent to which the interests of the “speaker” align with those of the “listener” and thereby test whether this had consequences for emergent organization; in particular, when the two were not aligned, would perception lead production or vice versa?

In achieving our goals, we are extremely limited by what natural-language data can tell us directly. Historical linguistics and sociolinguistics can shed a great deal of light on how established vowel systems change (Honeybone & Salmons, 2015), but can say little about how systems might have *emerged* in the first place, at least in spoken languages. Research on novel sign languages does allow us to observe the emergence of phonological systems (e.g., Sandler, Aronoff, Meir, & Padden, 2011) and these data have been very valuable. The birth of novel sign languages is rather sporadic and hard to predict, however, so it would be useful to complement such research with replicable laboratory experiments that can be scheduled in advance. Experimental approaches also allow us to control and manipulate variables that cannot be controlled or manipulated in natural contexts. A challenge of designing such experiments, however, is that we need to minimize as much as possible biases from participants’ linguistic experience. We need, in other words, to have them engage in a communicative task in which the communicative medium (a) is not one they are familiar with, but (b) shares relevant properties with natural language. If that can be achieved, it also gives us the freedom to manipulate variables of interest that are related to the nature of the medium.

4 A laboratory-language approach

Over the last decade and a half an experimental approach has been developed that answers our requirements. This approach, which we will term a laboratory-language approach (and was termed *Experimental Semiotics* by Galantucci, 2009), involves participants playing games in which they either learn a novel artificial language (e.g., Kirby, Cornish, & Smith, 2008; Sneller & Roberts, 2018) or collaboratively construct a novel communication system in the laboratory (e.g., Fay, Garrod, Roberts, & Swoboda, 2010; Galantucci, 2005; Roberts, Lewandowski, & Galantucci, 2015). The approach was devised primarily to investigate the emergence of language and of linguistic structure and can be distinguished from classic artificial-language learning approaches (e.g., Culbertson, Smolensky, & Legendre, 2012; Fedzechkina, Jaeger, & Newport, 2016; Hudson Kam & Newport, 2009) in the inclusion of a social component whereby participants are exposed to each other’s communicative output, either directly through interaction (e.g., Galantucci, 2005; Sneller & Roberts, 2018), or – in iterated learning experiments – indirectly through exposure in training to the output of previous participants (e.g., Kirby et al., 2008; Roberts & Fedzechkina, 2018).

A broad range of communication media have been employed in laboratory-language experiments, ranging from graphical or gestural communication (e.g., Bergmann, Dale, & Lupyan, 2013; Dale & Lupyan, 2010; Fay et al., 2010; Galantucci, 2005; Micklos, 2016) to auditory (Verhoef, Kirby, & de Boer, 2014) or even tactile (Trendafilov, Lemmelä, & Murray-Smith, 2010) signaling. Taking a laboratory-language approach has a number of advantages. Analogues to change that would take many years in natural language can be observed in laboratory languages over very short time periods; factors that cannot be manipulated outside the laboratory, or cannot be manipulated in natural languages even within the laboratory, can be manipulated in laboratory languages with relative ease. By investigating processes of change in non-linguistic communication systems, particularly using novel signaling media, researchers can reduce the influence of the participants' own languages (for further discussion of the advantages of this approach, see Galantucci & Roberts, 2012; Roberts, 2017).

Although few laboratory-language studies, to our knowledge, have investigated the *organization* of combinatorial units in phonological spaces (de Boer & Verhoef, 2012, is something of an exception), several have investigated the *emergence* of combinatorial units from continuous signals (e.g., Del Giudice, 2012; Little, Eryılmaz, & de Boer, 2017; Roberts & Galantucci, 2012; Roberts et al., 2015; Verhoef et al., 2014). Our study was influenced by these: Participants played a cooperative signaling game, moving their fingers in a continuous space to select discrete color signals to send to each other to communicate a set of animal referents. We manipulated the relationship between finger position and color and measured the organization of the resulting systems. This task had several desirable features for our purposes. First, it allowed us to observe the emergence of a quasi-phonological system from scratch. Second, the communication medium has features in common with natural-language – in particular, the production of discrete units from a structured continuous space in which perceptual units have a reliable but non-trivial relationship with the articulatory movements that produced them (and can be measured independently of them). Third, the medium is different enough from speech or sign that we can hope to have minimized the influence of natural language on participants' behavior. Fourth, the task allowed us to easily manipulate features of the communication medium (namely the relationship between articulation and perception). Fifth, participant behavior can be measured rather precisely and involves the production of well defined signals in a relatively low-dimensional space. Finally, the task allows the role of cultural evolutionary processes such as communicative interaction to be directly investigated.

5 Method

5.1 Participants

Sixty undergraduate students (34 female), none of whom suffered from color-blindness, participated in dyads for course credit.[‡]

5.2 Materials

Participants sat in separate cubicles, each with a computer (a mid-2014 Apple iMac with a 21.5" screen), running custom-designed software written in Python (Python Software

[‡] Owing to a software error, other demographic data such as age and handedness were not recorded. However, their distribution is not expected to have deviated substantially from that of the wider undergraduate population.

Foundation, www.python.org) and Kivy (Virbel, Hansen, & Lobunets, 2011; www.kivy.org), and a wireless multitouch trackpad (a 2009 Apple Magic Trackpad, measuring 13.01cm by 13.13cm). Participants could not see each other from their cubicles or hear each other easily.

5.3 Procedure

Pairs of participants played a cooperative communication game, taking turns to be *Sender* and *Receiver*.[§] Each participant (henceforth *player*) in a dyad sat in a separate cubicle and saw a screen divided vertically into two halves. (For the most part, the screen looked much the same whether the player was Sender or Receiver; Figs. 2 and 3). In the left half of the screen – the *referent panel* – a set of *referents* were displayed (black animal silhouettes, a subset of those used by Roberts & Galantucci, 2012; Fig. 4).[¶] The top right quarter of the screen, the *color panel*, appeared gray by default, but would change color depending on the behavior of the Sender. The same was true of a smaller section immediately below it – the *sent-color panel* – which was also gray by default and took up a quarter of the width of the screen as a whole and a quarter of the height. (See Section 5.4 below for a description of how the color panel and the sent-color panel worked.) To the right of the sent-color panel, a timer was displayed on a white background. Below this, taking up half the width of the screen, was a *score panel* displaying the dyad's joint score against a black background.

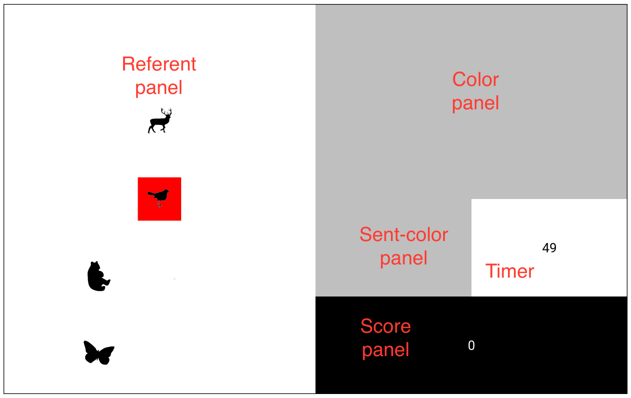

Figure 2: Sender's screen. Labels are for clarity and were not shown to participants.

The referent panel differed slightly for the Sender and the Receiver. First, the referents were not in the same places (i.e., were redistributed at random) from round to round. Second, no referent was ever in the center of the Receiver's referent panel; the Sender, on the other

[§] It was important for our question that both members of each dyad have an equal opportunity to be Sender and Receiver. Had this not been the case, any differences between conditions might be explicable in terms of a failure on the part of the Sender to appreciate the Receiver's needs. This approach also had the advantage of greater ecological validity.

[¶] Roberts and Galantucci (2012) used 20 referents in total; we used a 12-referent subset of theirs in order to give participants time to refine their signaling systems. Given the time available, continuing to add referents until there were 20 would have meant that systems would be in a constant state of flux.

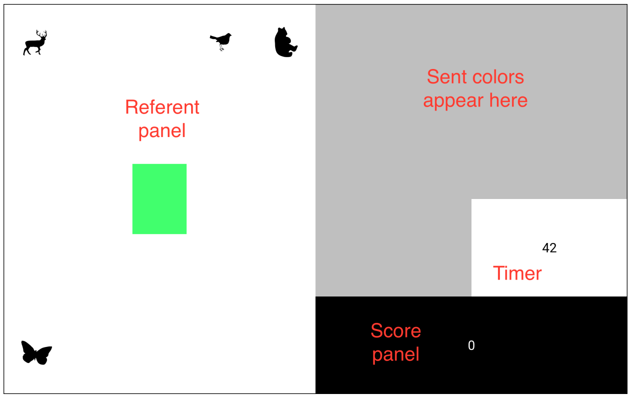

Figure 3: Receiver's screen. Labels are for clarity and were not shown to participants.

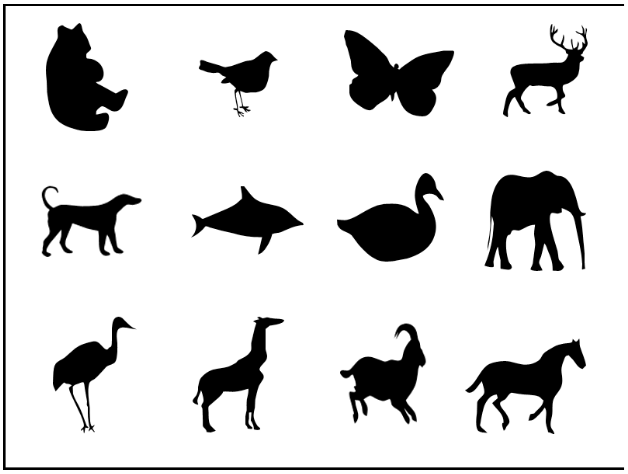

Figure 4: Referents used in the experiment. The top row appeared at the beginning of the game; as players became more successful, the other rows were added in turn.

hand, always had one referent in the center, against a red background (Fig. 2). This varied from round to round and was selected at random, by the server, from the set of available referents. Third, the Receiver had a green cursor that could be moved around the referent panel by using the arrow keys on the computer keyboard (Fig. 3); the Sender had no such movable cursor.

The Sender's task was to convey to the Receiver which referent was highlighted in the center of their referent panel by sending series of colors to the Receiver (see Section 5.4 below), and the Receiver's task was to move their cursor to the correct referent and press enter. Both players would then receive feedback: The correct referent would be highlighted in the referent space for the Receiver and the chosen referent would be highlighted for the Sender. This happened whether or not the Receiver chose correctly. If the Receiver did choose correctly, the dyad would score one point; their total point score was displayed throughout the game in the score panel at the bottom of the screen. After players started to do well at signaling the referents, more were added, in groups of four, up to a total of twelve. This would occur if, for all referents in the referent panel, the Receiver had selected them correctly at least 75% of the time over the previous four rounds in which they had occurred (cf. Roberts et al., 2015). Once referents were added, they were never removed and would continue to occur as targets, even if players started to do badly.

A round lasted 20s in total, with feedback lasting an additional 2s. If the Receiver had not chosen a referent by the time the 20s were up, the dyad scored no point for that round. Whatever the outcome of the round, the players would swap roles for the following round. The game lasted for 80min in total, and would finish at the end of the current round when the 80min mark had been passed. At the start of the experiment, players played four practice rounds that differed from the ordinary rounds in three ways: First, they lasted 60s rather than 20s; second, the players' score from these rounds did not carry over into the normal rounds; third, players were reminded at the start of each round whether they were Sender or Receiver. Beyond being told to move a finger around the pad and observe the screen, and to hold a finger down for 1s to send a color, players were not instructed how to use the signaling medium, but rather had to explore it on their own.

5.4 Signaling medium

To convey to the Receiver which referent to select, the Sender could send a series of colors. This could be achieved by moving one finger around on the trackpad, which would produce a color in the color panel on the top right of the Sender's (though not Receiver's) screen, which would change in real time depending on the coordinates of the Sender's finger. If the Sender took their finger off the pad or touched the pad with more than one finger, the color panel would appear gray. If the Sender held their finger in place on the trackpad for 1s or longer, the same color would appear for 2s both on the Receiver's color panel and on the Sender's sent-color panel. (This 2s period was fixed and was not influenced by how long the Sender held their finger down; in other words, duration was not a variable property of the color units.) This was the only means by which the Sender could send information to the Receiver.[‖] A Sender could send as many colors as they liked – including none at all – within the time available (20s).

The relationship between the Sender's finger position and the color produced was based on an RGB color space, with each color composed of a red, a green, and a blue component, the contribution of each ranging from 0 to 1 (e.g., the vector [1, 0, 0], where the digits indicate the red, green, and blue components respectively, would correspond to a bright red

[‖] In this respect, our study differs from earlier laboratory-language studies on combinatorial systems, which were mostly concerned with investigating the emergence of atomic units from continuous media. Participants in those studies were thus not provided with preordained means of producing units, with the consequence that identifying how such units might be constituted is itself a challenging task (Roberts & Galantucci, 2012). Because of this, and because we were concerned not with the emergence of such units, but how they become organized, our task forced subjects to select units from a continuous space, thereby simplifying our analysis while still retaining a continuous signal space from which units could be drawn.

color). The basic value for one of the three components increased from 0 to 1 as the Sender's finger moved from right to left on the pad, while another decreased from 1 to 0 in the same direction; the third color component increased as the finger moved vertically. Which color corresponded to which direction was counterbalanced between dyads, but for any given dyad, the exact center of the pad corresponded to the vector [0.5, 0.5, 0.5]. If vertical position corresponded to the blue component, then placing the finger in the middle of the top edge of the pad would produce an equal mixture of red and green [0.5, 0.5, 0], while the middle of the bottom edge would produce a mixture of red, green, and blue, with blue predominating: [0.5, 0.5, 1]. Players were not in fact exposed precisely to the basic color values described here; instead, the values were modified in a way that varied between two conditions. The details of this are described in Section 5.5.

5.5 Conditions

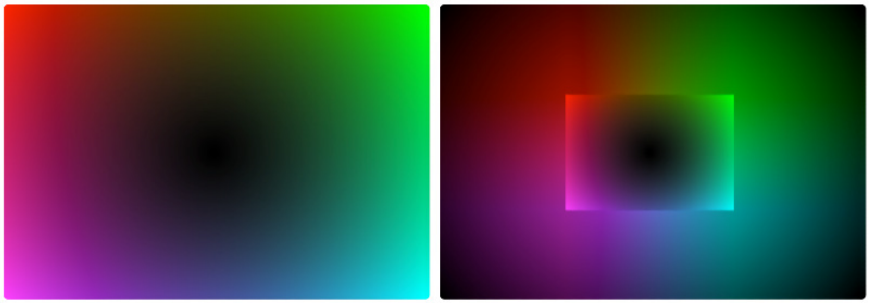

Figure 5: Example color spaces for Outer-edge and Inner-edge conditions respectively. Two points should be noted. First, participants never saw the space itself, only individual colors. Second, it is an artifact of this representation that colors drawn from the center area of both spaces appear more indistinguishably dark than they in fact were.

There were two conditions. In the *Outer-edge* condition the basic color values described above were altered depending on how close the Sender's finger was to the center of the pad (Fig. 5). This was done by multiplying the color component values by a modifier that ranged from 0 to 1. The modifier was calculated as d/d_{oe}, where d equals the Euclidean distance between the Sender's finger and the center of the space and d_{oe} equals the distance from the center of the space to the outer edge. This meant that colors towards the outer edges of the space were likely to be easier for the Receiver to distinguish. Since the edges of the pad were also easier to find reliably for the Sender, the pressures acting on the Sender and Receiver were therefore relatively aligned in this condition. Figure 6 shows an example of a "word" created in the Outer-edge condition (for a dyad whose color space was as in Fig. 5).

In the *Inner-edge* condition this was not the case. Here, an imaginary line was drawn 30% of the way in from the edge of the pad. Between the real edge of the pad and this "inner edge", the modifier was calculated as $1 - (d/d_{oe})$. Once the Sender's finger crossed the inner edge, however, the modifier changed to d/d_{ie}, where d_{ie} is the distance from the

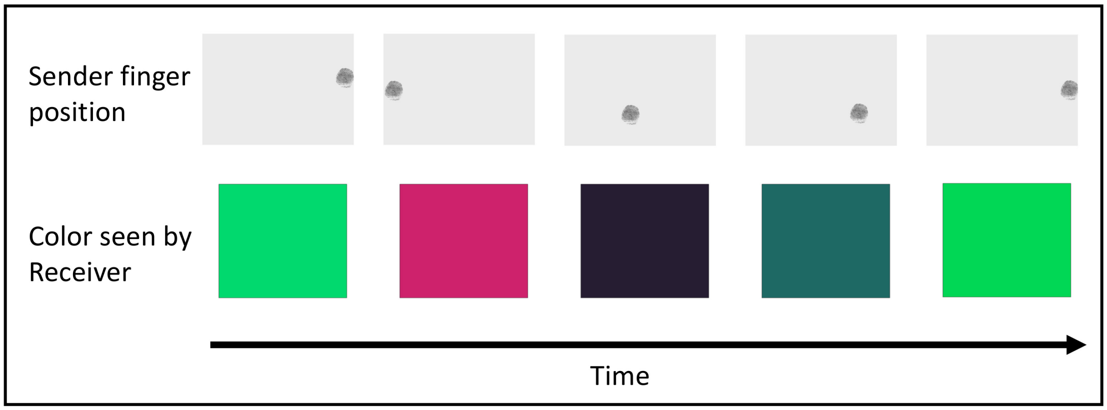

Figure 6: Example five-phoneme word taken from a trial in the Outer-edge condition. The Sender's finger position and the corresponding color sent to the Receiver are shown for each phoneme. (The fingerprint image used in this figure was made available on Wikimedia Commons by user Cyrillic, under the Creative Commons Attribution-Share Alike 3.0 Unported license. The original is available at https://commons.wikimedia.org/wiki/File:Fingerprintforcriminologystubs2.png.)

center of the pad to the inner edge. This meant that the colors got brighter as the Sender's finger moved away from the center of the pad, but then began abruptly to get darker again. The most convenient parts of the pad for the Sender to select reliably were still along the outer edge of the pad, but the easiest colors to distinguish for the Receiver were closer to the inner edge. The inner edge was in no way marked on the pad or screen; it became apparent to the Sender as they moved their finger around the pad and observed the effect.

5.6 Dependent variables

5.6.1 Identifying phonemes

For each player, we looked at referents for which the player could be determined to have established a signal, along with the signals they used for them. Establishing a signal meant that the referent had been successfully communicated at least three of the last four times it had occurred (which was also the criterion used to decide, during the game, whether more referents should be added). This gave us a set of "words" which that player had in their system, each of which was composed of a set of *units* – colors that players had chosen to send. Each unit existed in both a production space and a perceptual space and thus consisted of two sets of coordinates. First, it consisted of an x and a y coordinate corresponding to the Sender's finger position when the color was sent. Second, it consisted of the RGB coordinates of the color displayed to both players on the screen. Within each space, coordinates could be compared across conditions and, given knowledge of what condition the unit came from, the RGB coordinates could be inferred deterministically from the xy coordinates. Because two of the dimensions in the RGB space were mapped to the same (horizontal) dimension on the trackpad, two of the colors could never vary independently. Therefore, while we will discuss phonemes in the perceptual space as three-dimensional, the primary analysis was performed on the two-dimensional articulatory space.

For each player, we pooled all the units across the words they had established to produce a “phoneme inventory”. As for natural language, it was assumed that players might reuse phonemes between words, or within words. However, the potential phonemes were drawn from a continuous space, so the precise coordinates of any phoneme might differ slightly across realizations, making it non-obvious whether two similar but non-identical units (such as the first unit and last unit in Fig. 6) are the same phoneme or not. The same issue applies, of course, to natural language. The /p/ at the start of *peel*, for instance, is not phonetically identical to the /p/ in *spam*. Identifying what phonetic units correspond to the same phoneme thus poses a challenge for linguists studying a new language. Indeed, it can pose a challenge for linguists investigating change and variation in well studied languages too, most obviously in attempting to identify if two sounds have merged or remain separate for a given individual (as with *pin* and *pen* for many American English speakers). Nycz and Hall-Lew (2013) reviewed four statistical approaches to this task. The simplest method is to measure the Euclidean distance between vowels, which necessitates selecting a somewhat arbitrary cutoff below which vowels are considered to be merged (cf. Roberts et al., 2015). Another method uses mixed effects regression to take into account fixed effects such as phonological environment or WORD CLASS, with WORD as a random effect. This allows differences between F_1 and F_2 to be estimated, which can be used to calculate an adjusted Euclidean distance measure. It also produces a measure of significance for the WORD CLASS variable, though not of the Euclidean distance itself. Another approach involves using spectral overlap, whereby ellipses are generated and an overlap fraction generated. Like the other measures, however, no measure of significance is generated. The fourth option (introduced by Hay, Warren, & Drager, 2006, and probably now the preferred approach in measuring whether two vowels have merged) is to calculate a “Pillai score” (formally, a Pillai-Bartlett trace), a statistic which represents the proportion of one variance that can be predicted by another variance and which is generated as part of a MANOVA. The Pillai score ranges from 0 to 1, where a higher number indicates a greater difference between distributions (see DasGupta, 2005 for a formal account and Hall-Lew, 2010 for a description of how to calculate it in R). It also produces a measure of significance, allowing a cutoff to be established in a somewhat less arbitrary way from the other measures discussed. Given its popularity, its convenience, and its suitability for our data, we used the Pillai method to identify phonemes in our data. This was done as follows.

First, because the measure relies on having several instances of the same word available, we used every successful word that a player had produced (as opposed to, say, the last successful word only). Because different words for the same referent would sometimes have different lengths, we calculated the mode length for that referent and discarded any words of a different length. We assumed that for a given referent all the different color units used for each word position (e.g., first color unit in the word, second color, third color etc.) could be treated as instances of the same phoneme and that this phoneme, represented as a distribution of those instances, could thus be compared with other phonemes from the same and from different words. We then calculated Pillai scores for all pairs based on their x and y coordinates in the articulatory space. If the p-value for the pair was greater than or equal to 0.001, we treated them as instances of the same phoneme. In such cases we also calculated the mean Euclidean distance between the pair. These distances were used to calculate an overall mean that would serve as a *mean Pillai cutoff.* We needed this because, for some referents for some players, too few successful words were generated to calculate a Pillai score. For these we judged two units to constitute the same phoneme if the Euclidean distance between them was below the mean Pillai cutoff. This resulted in a set of "phoneme clouds", each cloud being a collection units that had been judged to be instances of the same phoneme. In forming such clouds, a difficulty presents itself: How to deal with a case where three potential phonemes (A, B, and C) are being compared, and B overlaps sufficiently with A and C to be considered the same phoneme as each of them, but A and C behave as clearly distinct from each other? In principle (though this did not occur) all phonemes in the data could turn out to form one long phoneme continuum. In such cases, should the system be judged to consist of one single phoneme or a set of overlapping ones? We decided on the latter option. For the A, B, C example just given, we would treat the three clouds as representing two phonemes. (In the event, this did not make a difference to our pattern of results.)

We then measured the following variables. In each case, the mean value for a dyad forms the basis of the results reported below.

Number of referents. The number of referents for which a word had successfully been established.

Success index. This measure was based on how many words a player successfully established and how fast they did so, giving us a more fine-grained measure of success than number of referents alone. For every round of a given game, we counted how many referents each player had an established word for (see above) at that point. We then calculated a success index as $(\sum_1^{n_r} s)/12n_r$, where n_r is the number of rounds and the numerator is a cumulative count of s, the number of successfully established words in a given round (with 12 being the maximum possible given the number of referents).[**]

Word length. The number of phonemes in each word.

Phoneme inventory size. The number of phonemes in an inventory.

Dispersion. Three measures of dispersion were used:

Distance from center. The mean Euclidean distance between the center of the space and the centers of the phoneme clouds. This was then divided by the distance from the center of the space to its corner (i.e., the maximum possible distance) to give a number between 0 and 1.

[**] It should be noted that no player could actually score 1, as that would require them to have successfully communicated all twelve referents several times before the start of the game. While this could be accounted for, this would complicate the calculation, which we did not deem necessary for a relative measure of success.

Mode brightness. Because of the nature of the mapping between finger position and colors, distance from the center of the articulatory space would not necessarily correspond to dispersion in the perceptual space. As an equivalent measure in RGB space to the distance-from-center measure in the articulatory space, we took the mean, for all colors, of the distance of the brightest component from 0. As the color components were recorded on a scale of 0 to 1, this corresponds simply to the value of the brightest component.

Pairwise distance. A system in which all the phonemes were clustered in one corner, or along one edge, would score relatively highly in terms of distance from center, but would not be very well dispersed. For comparison with distance from center in the articulatory space, we therefore measured the mean pairwise Euclidean distance between the centers (i.e., the mean) of all the phoneme clouds. (Sets with fewer than two phonemes would have been excluded from this measure, but there were no such sets.) This was then divided by the maximum possible mean distance for a set of that size (see Appendix B) to give a number between 0 and 1.

We present comparisons between conditions for all variables below. This is based in every case on a randomization test in which we shuffled the data between the two conditions 100,000 times and counted how many times a difference at least as large as the real difference occurred by chance, dividing the result by 100,000 to yield a p-value (for further discussion of randomization tests, a form of permutation test, see Edgington & Onghena, 2007). We compared results with chance using a similar method, in which we generated 100,000 random phoneme systems (Section 6.6). We also examined *iconicity*. For any referential communication task in which participants must communicate using novel signs, it is likely that they will attempt to come up with signals that have a non-arbitrary relationship with their referents (cf. Roberts et al., 2015, who found in another study on combinatoriality that participants preferred iconic strategies where possible; see also Caldwell & Smith, 2012; Fay et al., 2010). Even though the referents we used were deliberately monochrome, participants might still do such things as disproportionately choose brown to represent bears, gray to represent elephants, or blue to represent dolphins. It is important to note that this should not have varied between conditions, so should not represent a confound. Identifying iconicity, which is to a great extent subjective, is non-trivial (for a discussion of the issues, see Roberts & Galantucci, 2012). For our purposes we took, for each final "word" used by each dyad, the mean R, G, and B values and tested whether any of these values were predictable across dyads based on referent. The logic here is that, if there is an effect of iconicity, we should see a preference for certain colors across dyads for particular referents.

6 Results

Fig. 7 gives examples of phoneme inventories, including the most and least dispersed sets. Data were parsed using Python; analyses were conducted using the R Statistical environment (R Core Team, 2014); linear models were run using the lme4 and lmerTest libraries (Bates,

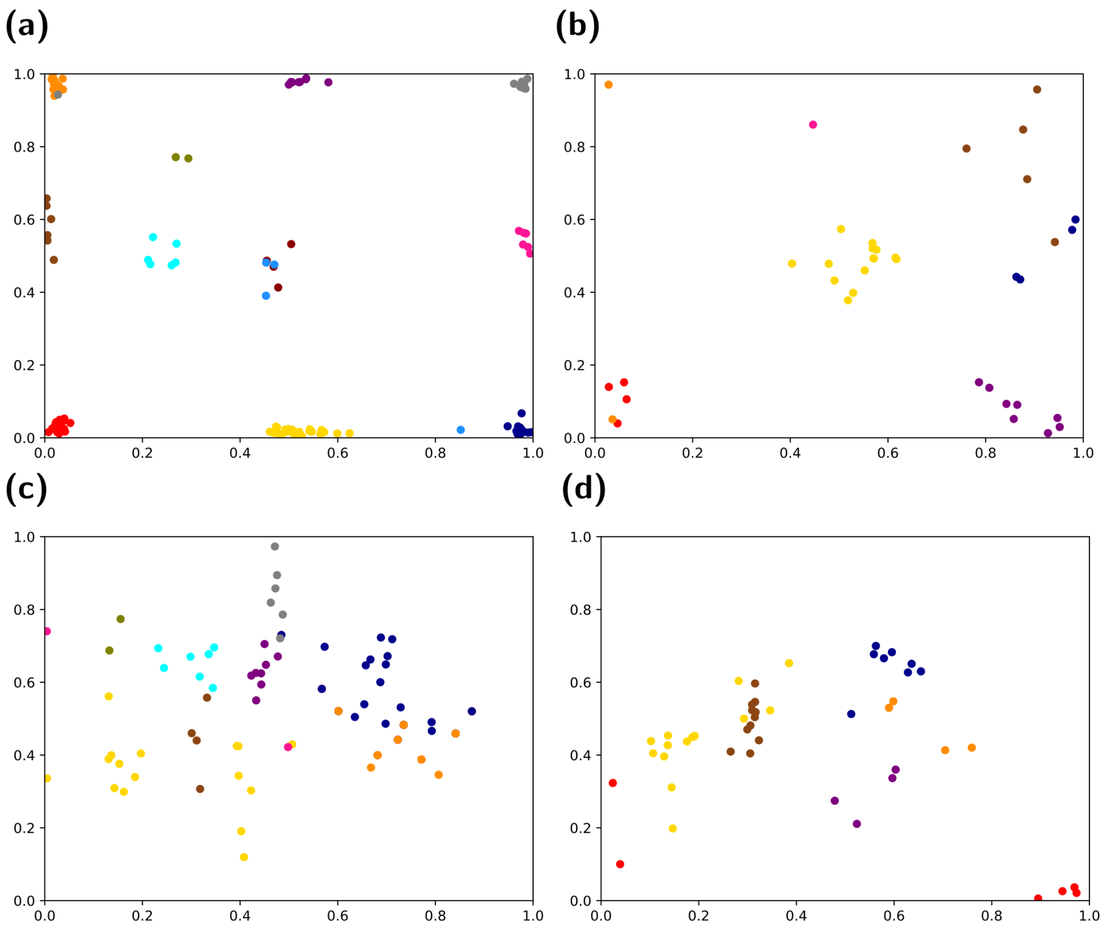

Figure 7: Example phoneme inventories, with different colors corresponding to different phonemes: (a) inventory with highest pairwise distance, (b) inventory with highest distance from center, (c) inventory with lowest pairwise distance, (d) inventory with lowest distance from center. Sets a and b were both found in the Outer-edge condition while sets c and d were both found in the Inner-edge condition. Each point indicates a single instance of a phoneme being used; phonemes are distinguished by color, but this does not reflect the corresponding color in RGB space (which would vary across an individual phoneme). Axes simply indicate normalized xy coordinates. In each case the inventory of one member of the dyad in question is shown, for simplicity's sake. In the analyses presented in the text, the mean values for each dyad were used.

Mächler, Bolker, & Walker, 2015; Kuznetsova, Brockhoff, & Christensen, 2017), and plots were created using the ggplot2 and psych libraries (Revelle, 2018; Wickham, 2016). Data and analysis scripts are available at osf.io/3c4zb/.

6.1 Iconicity

To test for an iconicity effect we fit three mixed effects models with mean color value as the outcome variable in each case, referent as predictor, condition as an interaction term, and dyad as a random factor. There were effects for all colors and for all referents except elephants, but there was no main effect of or interaction with condition, consistent with the expectation that this would not present a confound. The full results of the model are presented in Appendix A.

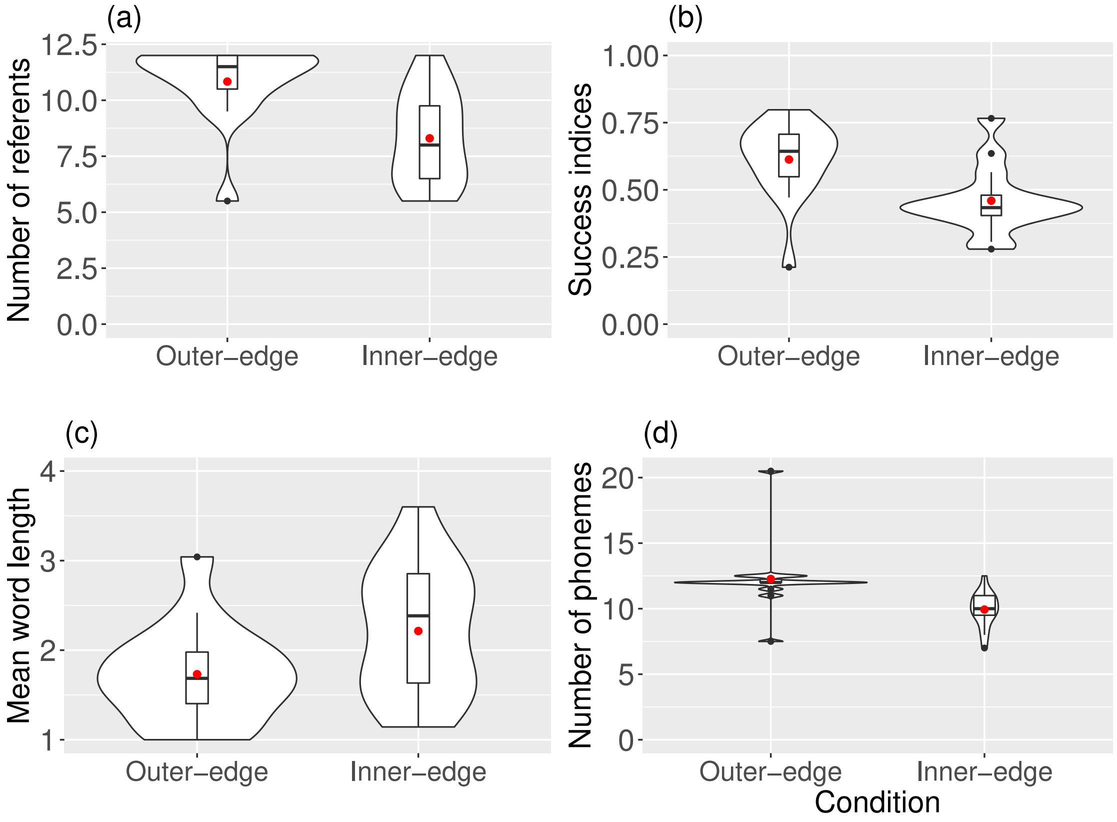

Figure 8: Violin plots showing (a) number of referents, (b) success indices, (c) mean word length, and (d) number of phonemes for each condition. Red dots indicate means; black dots indicate outliers.

6.2 Number of referents and success

Players in the Inner-edge condition found the game harder (Figs. 8a and b), consistent with our expectation that breaking the alignment between production ease and perceptual ease would pose a communication challenge. They established successful signals for a mean of 8.3 referents out of 12 (SD = 2.04), and their mean success index was 0.46 (SD = 0.12). Only one dyad in this condition succeeded in establishing signals for all twelve referents, and only seven dyads (46.7%) established signals for more than eight referents (though every dyad at least got beyond the first set of four). Dyads in the Outer-edge condition did significantly better in each case with a mean of 10.83 referents (SD = 1.72; $p = 0.0015$, mean difference = 2.53) and a mean success index of 0.61 (SD = 0.14; $p = 0.004$, mean difference = 0.15).

6.3 Word length

Words were slightly shorter in the Outer-edge condition than the Inner-edge condition (Fig. 8c), with a mean length of 1.7 phonemes in the former (SD = 0.54) and 2.2 (SD = 0.72) in the latter: $p = 0.049$ (mean difference 0.5). This may reflect greater difficulty in the Inner-edge condition of selecting distinctive colors, necessitating more combination of colors. Indeed, as discussed in Section 6.4, phoneme inventories were larger in the Outer-edge condition.

6.4 Phoneme inventory size

Phoneme inventories were smaller in the Inner-edge condition than the Outer-edge condition (Fig. 8d): $p = 0.0009$ (mean difference 2.33 phonemes). This is consistent with the evidence that participants found it harder to construct systems in this condition. First, the lesser alignment between production ease and perceptual ease meant that it was harder to establish stable phonemes; second, as a result of this, participants in this condition had fewer stable words, so there was less pressure to increase the phoneme inventory.

6.5 Dispersion

Regardless of how dispersion was measured, phonemes were more dispersed in the Outer-edge condition (Fig. 9; see Table 1 for comparison statistics). This supports the view that Senders took perceptual information into account in choosing phonemes to send (as opposed to selecting colors without regard for ease of perception, which would lead to identical results between conditions). The two dispersion measures were well correlated: $r(28) = 0.97$, $p = 2.2 \times 10^{-16}$ (Fig. 10). However, distance from center was significantly lower than pairwise distance, $p = 8 \times 10^{-05}$ (mean difference 0.11). The same was also true for each condition individually, although the difference was considerably greater in the Outer-edge condition, $p < 0.001$[††] (mean difference 0.14), than in the Inner-edge condition, $p = 0.049$ (mean difference 0.08).

The obvious next question is whether levels of dispersion were more similar between conditions in the perceptual RGB space. If participants were maximizing perceptual dispersion to a similar degree across conditions, this would inevitably lead to different levels of production-space dispersion between conditions. Surprisingly this was not the case. As can be seen in Fig. 9 and Table 1, the difference between conditions in the production space was repeated in the perceptual space. This is likely to be another consequence of the misalignment between ease of production and ease of perception in the Inner-edge condition. While a similar level of perceptual dispersion was *available* to participants in the two conditions, it was harder to achieve it in a stable way, as it relied on participants being able to locate the same color reliably – much easier when that color is located on the edge of production space.

[††] Where specific p-values are not given, as here, this indicates that a value equal to or greater than the true value did not occur in any of the 100,000 permutations of the randomization test.

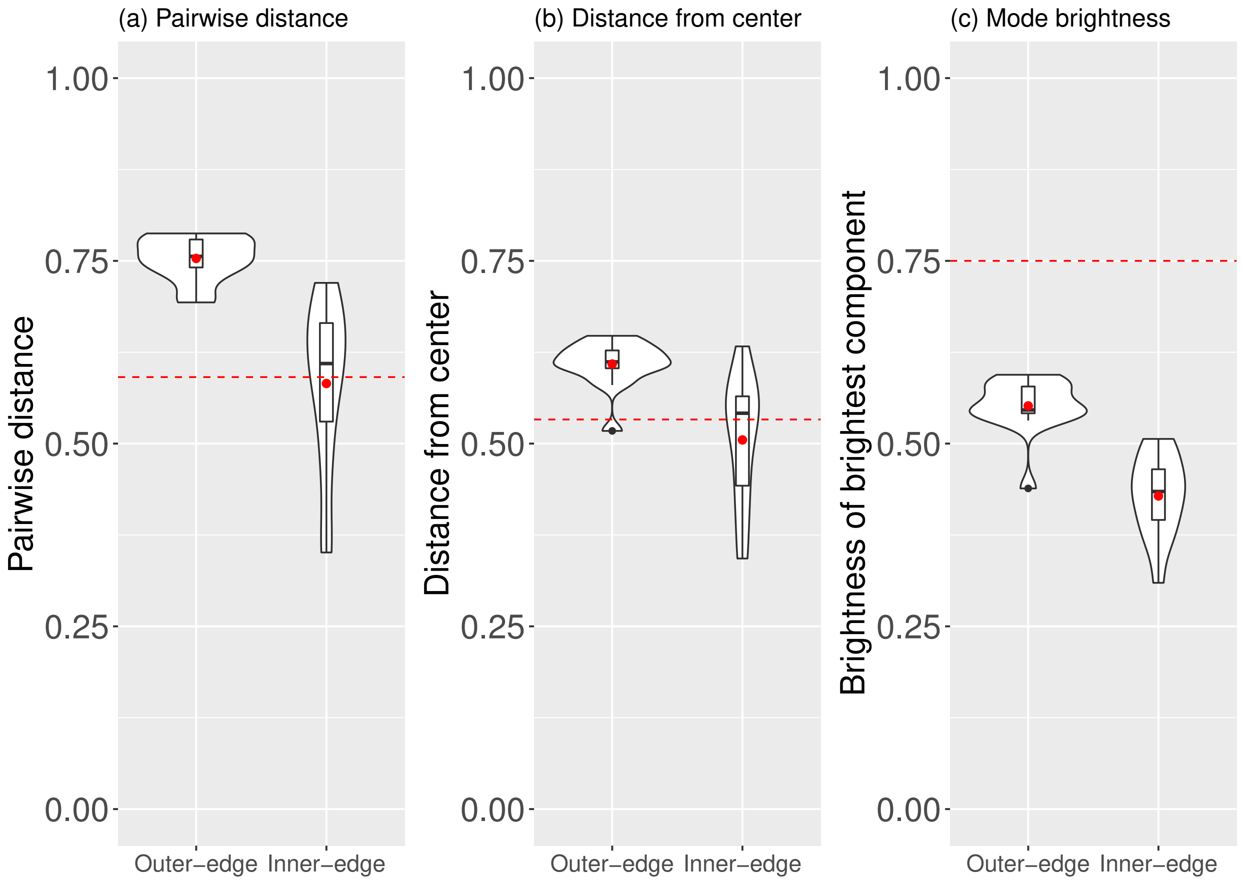

Figure 9: Violin plots of dispersion patterns: (a) Pairwise distance (xy space), (b) distance from center (xy space), and (c) mean brightness (RGB space). Red dots indicate means; black dots indicate outliers. Dotted line indicates chance-level dispersion.

Faced with this, it seems that participants in the Inner-edge condition were more likely to make do with less bright colors.

Table 1: Comparison between conditions for dispersion measures

	mean diff.	p
Pairwise distance	0.17	< 0.001
Distance from center	0.1	4×10^{-05}
Mode brightness	0.12	< 0.001

Table 2: Table of p-values based on comparison of real inventories with random inventories. Given the randomization, all values are approximate.

Condition	Pairwise distance	Distance from center	Mode Brightness
Outer-edge	< 0.0001	7×10^{-5}	1
Inner-edge	0.051	0.87	1

6.6 Comparison with chance

We compared the results with chance by generating, for both the production space and the perceptual space, 100,000 random phoneme inventories, matching the distributions of phoneme inventory sizes actually produced in each condition. These random inventories were constrained to ensure that no two phonemes could be closer than the minimum distance between phonemes in any actual inventory from the same condition. For each random inventory we measured dispersion and symmetry as for the real sets. For each measure, set, and condition, the mean value for the 100,000 random inventories was used as a measure of what level of dispersion of symmetry we could expect by chance. This is indicated by a red-dotted line in Fig. 9. We also used this method to calculate p-values. In particular, we counted every time a particular measure on a random inventory produced a value equal to or higher than the value calculated from the corresponding real inventory. We then divided this tally by 100,000 to generate a p-value. These are given in Table 2. The main finding is that dispersion in the production space (regardless of measure) was significantly greater than chance in the Outer-edge condition, but not in the Inner-edge condition. These results are consistent with the view that perceptual demands were shaping the "articulatory" decisions of participants. If participants cared about clearly distinct colors in both conditions, this would lead to significant dispersion in the production space in the Outer-edge condition (where the outer edges of the space were associated with brighter colors) and for phonemes to be much closer to the center in the Inner-edge condition. Mode brightness was not greater than chance for any condition, suggesting that brightness per se, as opposed to perceptibility, was not driving behavior.

6.7 Relationships between variables

Fig. 10 shows relationships between the different dispersion measures as well as relationships between them and success. All variables were significantly correlated, with a particularly striking relationship between pairwise distance and distance from center. This relationship also holds for each of the conditions individually, suggesting that the two measures are capturing the same thing: $r(13) = 0.98$, $p = 8.963 \times 10^{-11}$, for the Inner-edge condition and $r(13) = 0.81$, $p = 0.0003$ for the Outer-edge condition. (The same is also true for mode brightness, with $p < 0.01$ in every case.) However, the relationships with success do not hold for any condition considered individually. This suggests that success was driven by

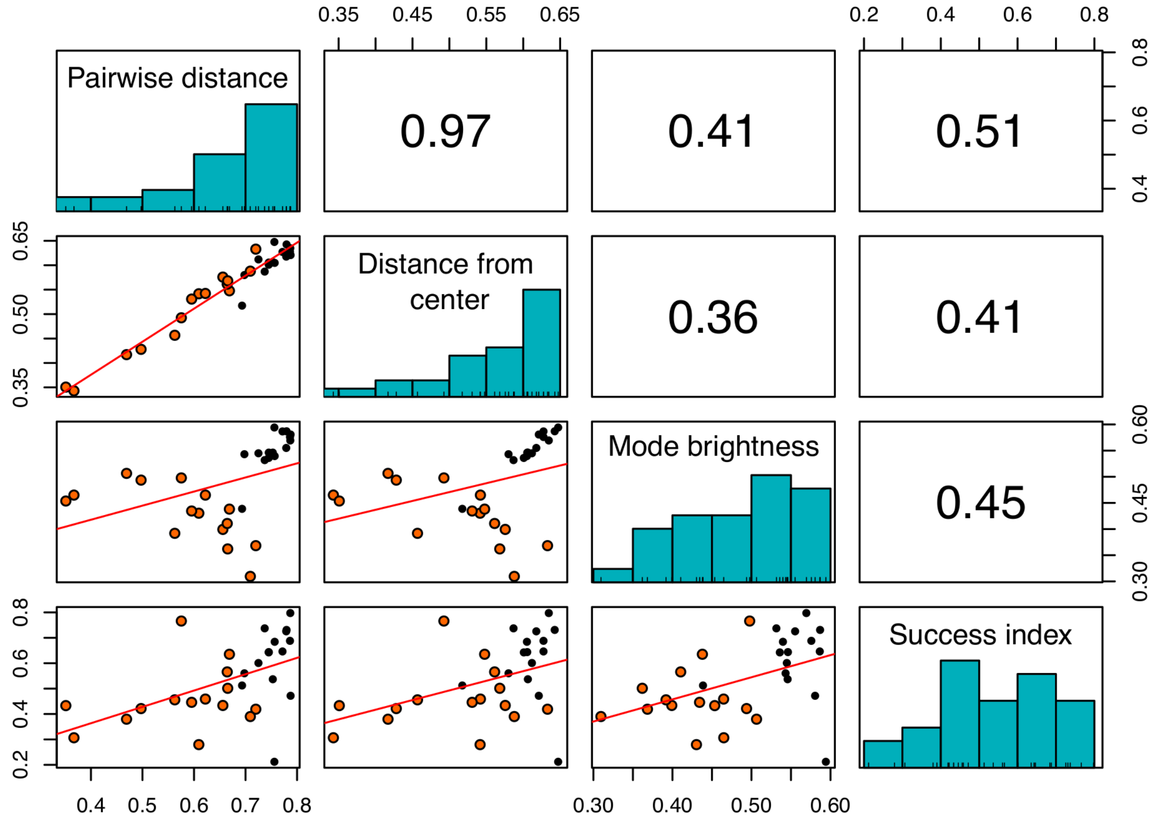

Figure 10: Matrix of relationships between dispersion measures and success index. Orange dots are from the Inner-edge condition and blue dots from the Outer-edge condition. Numbers in upper plots are Pearson's R values.

condition (i.e., the alignment of the perceptual space with the production space) rather than by dispersion.

7 Discussion

We set out with two goals. The first was to establish a new experimental approach to investigating the cultural-evolutionary processes involved in phonological organization. Our results suggest that our approach is promising: The task was achievable by participants; we were able to analyze the data as we might analyze natural-language data; the resulting quasi-phonological systems should at least a qualitative resemblance to natural-language systems in many respects; and we were able to draw conclusions about how such systems become organized.

This last point brings us to our second goal, which was to answer three specific questions about the organization of combinatorial units in our participants' communication systems. The first was whether the visual communication systems that evolved would exhibit above-chance levels of dispersion. For the production space, this was confirmed for dispersion measures in the Outer-edge condition, but was not the case for any measure in the Inner-edge condition, or for mode brightness in the perceptual space. The second question was whether our measures of dispersion would correlate with communicative success. We found that, while this appeared to be the case for the data as a whole, the relationship did not hold for any condition considered individually, suggesting that success was driven by condition (i.e., by the alignment between perception and production demands) and not by dispersion. This leads us to our third question, which concerned the relationship between sender and receiver: If the pressures acting on one were misaligned with the pressures acting on the other (as in our Inner-edge condition), which set of pressures would win out? The results suggest that the perceptibility demands of receivers were allowed to guide the behavior of senders. This is evidenced by the differences in dispersal between the two conditions in the production space (Fig. 9) and suggests that the high levels of dispersion observed in natural language vowel spaces (and, indeed, other phonological systems) owe something to inherent alignment between the articulatory space and the perceptual space.

Such alignment supports communication by helping language users produce distinct phonological units reliably. The difference between the conditions of our experiments in terms of success rate is good evidence that the misalignment in the Inner-edge condition made it harder for senders to satisfy perceptibility demands reliably. There is also evidence from both conditions, based on the below-chance levels of mode brightness, that participants were aiming for discriminability rather than brightness in the perceptual space, which is functionally preferable for communication.

At this point a potential limitation in our design should be noted. In order to create a space in which production and perception were misaligned, it was necessary to create a space in which communication was more difficult than when the two are well aligned. This seems to be an inevitable consequence of misalignment. However, it is possible that our manipulation inadvertently made communication in the Inner-edge condition harder in ways that are not strictly about misalignment. In future work it would be useful to investigate different means of creating misalignment.

A second potential limitation concerns the mapping between the two spaces. Because the production space was two-dimensional and the perceptual space three-dimensional, two of the dimensions of the latter could not vary independently. In future work it would be interesting to increase the number of dimensions in the production space. A related issue concerns the use of RGB space rather than, for instance, CIELAB space in which the perceptual distance between colors is more uniform. The purpose of our experiment was to map finger position on to a continuous space in a reliable way; the precise nature of that space was not per se important, and RGB space was chosen because it allowed the correspondence between finger position and color to be implemented in a way that was particularly straightforward and transparent for participants. However, future work might benefit from the qualities afforded by more sophisticated color spaces. It would also be interesting to consider perceptual domains other than color. Any perceptual domain familiar to humans brings with it pre-existing cultural associations; linguistic color terms, for instance, are known to influence human perception of colors (Forder & Lupyan, 2019). By conducting further experiments employing different perceptual domains, we would be able to distinguish more clearly the role of domain-general processes from domain-specific ones.

A third opportunity for developing the paradigm in future work concerns population size. Communicating dyads allowed us to focus in on the dynamics of sender-receiver interaction, but real phonologies emerge in much larger populations (see Fay & Ellison, 2013 for an example of a visual laboratory-language task comparing dyads and larger populations). Similarly, our game lasted for a little over an hour, which is long compared with many experiments but is incredibly short compared with the periods over which natural languages evolve. While laboratory-language experiments allow language change to be observed extremely rapidly, the question of when to end trials remains somewhat open, and longer games with larger populations should be investigated in later work. (See Morin et al., 2018 for recent work in which very large numbers of participants engage in multiple brief interactions over long periods.) Including an operationalization of generational transmission would also contribute to understanding the longer-scale dynamics of our systems (Kirby et al., 2014). Furthermore, phonological systems should not necessarily be assumed to have arisen initially in populations with fully modern articulatory systems. It is likely, at least for speech, that there was a period of coadaptation between culturally transmitted phonologies and genetically transmitted vocal tracts (cf. de Boer, 2016). This is clearly not represented in our experiment, but we note that our paradigm could be modified to incorporate proxies for coadaptation by, for example, changing the relationship between the production and perceptual spaces between trials (or even within a trial) in response to participant behavior.

In spite of these potential limitations we were encouraged by the emergence of impressionistically vowel-like systems (see, e.g., Fig. 7) and above-chance levels of dispersal, as well as the results concerning the effect of alignment and misalignment of perception–production pressures. These results suggest that we should expect phonologies to adapt to regions of the phonetic space in which articulatory constraints and perceptual constraints are mutually reinforcing, consistent with accounts such as quantal theory (Stevens & Keyser, 2010), the distinctive region model (Carré et al., 2017), and dispersion-focalization theory (Schwartz, Abry, Boë, Ménard, & Vallée, 2005), in which the topology of the signaling space plays an important role.

Our results also support the view that communicative interaction may be a key mechanism for the emergence of communicatively adaptive signaling systems. Our participants constructed systems collaboratively; they did not learn them. It is therefore interesting to consider interaction in comparison with generational transmission. It has been claimed by several researchers that the cultural evolutionary emergence of structure requires a "bottleneck". This is most easily understood in the context of acquisition: Because a learner cannot be exposed during acquisition to every utterance they might want to produce, they are forced to generalize, leading to increasing levels of systematicity over generations (e.g.

Kirby, Tamariz, Cornish, & Smith, 2015). It is less obvious what constitutes a bottleneck in communicative tasks that do not involve learning, but a number of experimental studies over the last decade or so have shown that structure can indeed emerge in such circumstances (Nölle, Staib, Fusaroli, & Tylén, 2018; Raviv, Meyer, & Lev-Ari, 2019; Roberts & Galantucci, 2012; Roberts et al., 2015). One possibility, explicitly discussed by Raviv et al. (2019), is that bottlenecks can exist in interaction too, with a pressure for generalization arising from (e.g.) continuous communicative pressures to make new distinctions in the meaning space. This is consistent with the results of Nölle et al. (2018) in which a relationship was observed between the emergence of structure and the openness of the meaning space. However, Raviv et al. (2019) did not in fact find such a relationship in their own data, while Roberts and Galantucci (2012) found only a weak relationship between the number of referents expressed and levels of structure. Roberts et al. (2015) found no direct evidence of a relationship although, in their study as in ours, referents were added over the course of the experiment, and it is possible that this openness in the meaning space played a role in driving the emergence of structure overall, as in the work by Nölle et al. (2018). In future work, it would be interesting to compare the role this factor across different paradigms and kinds of structure.

An alternative possibility, however, is that a bottleneck is in fact not required for all kinds of structure. Researchers such as MacNeilage and Davis (2000) and Studdert-Kennedy and Goldstein (2003) have argued, for instance, that the units of spoken phonology are ultimately grounded in "phylogenetically ancient mammalian oral capacities for sucking, licking, swallowing and chewing" (Studdert-Kennedy & Goldstein, 2003, p. 239). A way of putting their account is that a relatively small set of stable motor actions act as attractors in the phonetic phase space, so that simply repeating the same phonetic actions over and over, assuming some degree of noise, will tend to shift them towards such attractors, leading ultimately to a system structured by a relatively small set of parameters. A similar story could be told of the emergence of dispersal in vowel spaces, with stable points in the space – including points where the pressures acting on speakers and listeners are mutually reinforcing – acting as attractors. Such an account could of course be recast in terms of a bottleneck, though not one based straightforwardly on memory. In that case, it would be worth considering in future work how different linguistic spaces (in particular, combinatorial vs. compositional) might impose different kinds of bottlenecks and how they interact. How might the levels and kinds of structure we observed vary if repetition and interaction within dyads were reduced but generational transmission increased?

One other point to discuss concerns our method for identifying "phonemes" in our data. We used the Pillai method, which is already widely used for natural language research and has been argued explicitly to be preferable to certain alternative measures, such as simple Euclidean-distance-based options (Hay et al., 2006; Nycz & Hall-Lew, 2013). There was no space in this current study to undertake our own comparison of methods; however, we consider that there is a gap in the literature for a paper comparing alternative methods for identifying combinatorial structure in laboratory-language data.

In summary, we consider that the study we have presented strongly supports the view that a laboratory approach, and a cultural evolutionary approach, can shed useful light on the organization of phonological spaces in language (cf. Blevins, 2004; Vaux & Samuels, 2015). In particular, it suggests (in line with a number of theoretical models) that a significant portion of structure in phonological spaces is the consequence of interactive processes and sender-receiver dynamics, with perception to some extent guiding production, and the most stable systems arising where perception and production are mutually reinforcing. To put it another way, this study provides evidence that – as in so many evolving systems – the form of language reflects, at least in part, its function (Thompson, 1942).

Acknowledgements

We thank Mark Liberman, Lacey Wade, Yosiane White, Colin Twomey, and Chris Ahern for insightful comments, and we thank various members of the Cultural Evolution of Language Lab for running the experimental trials.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Bergmann, T., Dale, R., & Lupyan, G. (2013). The impact of communicative constraints on the emergence of a graphical communication system. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 35, pp. 1887–1892).

Blevins, J. (2004). *Evolutionary Phonology: The Emergence of Sound Patterns.* Cambridge: Cambridge University Press.

Brentari, D. (2011). Sign language phonology. In J. Goldsmith, J. Riggle, & A. C. L. Yu (Eds.), *The Handbook of Phonological Theory* (pp. 691–721). Malden, MA: Wiley-Blackwell.

Caldwell, C. A., & Smith, K. (2012). Cultural evolution and perpetuation of arbitrary communicative conventions in experimental microsocieties. *PLOS ONE*, *7*(8), e43807. doi: 10.1371/journal.pone.0043807

Carré, R., Divenyi, P., & Mrayati, M. (2017). *Speech: A Dynamic Process.* Berlin/Boston: de Gruyter.

Chomsky, N., & Halle, M. (1968). *The Sound Patterns of English.* Cambridge, MA: MIT Press.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, *122*, 306–329. doi: 10.1016/j.cognition.2011.10.017

Dale, R., & Lupyan, G. (2010). Squiggle: Large-scale social emergence of simple symbols. In K. Smith, A. Smith, M. Schouwstra, & B. de Boer (Eds.), *Proceedings of the 8th Evolution of Language Conference* (pp. 391–392). London: World Scientific. doi: 10.1142/9789814295222_0060

DasGupta, S. (2005). Pillai's Trace Test. In P. Armitage & T. Colton (Eds.), *Encyclopedia of Biostatistics* (Vol. 5). New York: Wiley. doi: 10.1002/0470011815.b2a13067

de Boer, B. (2000). Self-organization in vowel systems. *Journal of Phonetics*, *28*(4), 441–465. doi: 10.1006/jpho.2000.0125

de Boer, B. (2001). *The Origins of Vowel Systems.* Oxford: Oxford University Press.

de Boer, B. (2016). Modeling co-evolution of speech and biology. *Topics in Cognitive Science*, *8*(2), 459–468. doi: 10.1111/tops.12191

de Boer, B., & Verhoef, T. (2012). Language dynamics in structured form and meaning spaces. *Advances in Complex Systems*, *15*(3–4), 1150021 (20 pages). doi: 10.1142/S0219525911500214

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. doi: 10.1515/langcog-2012-0020

Edgington, E. S., & Onghena, P. (2007). *Randomization Tests: Fourth Edition.* Boca Raton, FL: Chapman & Hall.

Fay, N., & Ellison, T. M. (2013). The cultural evolution of human communication systems in different sized populations: Usability trumps learnability. *PLOS ONE*, *8*(8), e71781.

doi: 10.1371/journal.pone.0071781

Fay, N., Garrod, S., Roberts, L., & Swoboda, N. (2010). The interactive evolution of human communication systems. *Cognitive Science*, 1–36. doi: 10.1111/j.1551-6709.2009.01090.x

Fedzechkina, M., Jaeger, T. F., & Newport, E. (2016). Balancing effort and information during language acquisition: Evidence from word order and case marking. *Cognitive Science*. doi: 10.1111/cogs.12346

Fernández, E., Kalcsics, J., & Nickel, S. (2013). The maximum dispersion problem. *Omega*, *41*(4), 721–730. doi: 10.1016/j.omega.2012.09.005

Flemming, E. (2001). Scalar and categorical phenomena in a unified model of phonetics and phonology. *Phonology*, *18*, 7–44. doi: 10.1017/S0952675701004006

Forder, L., & Lupyan, G. (2019). Hearing words changes color perception: Facilitation of color discrimination by verbal and visual cues. *Journal of Experimental Psychology: General*, *148*(7), 1105. doi: 10.1037/xge0000560

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science*, *29*(5), 737–67. doi: 10.1207/s15516709cog0000_34

Galantucci, B. (2009). Experimental Semiotics: A new approach for studying communication as a form of joint action. *Topics in Cognitive Science*, *1*(2), 393–410. doi: 10.1111/j.1756-8765.2009.01027.x

Galantucci, B., & Roberts, G. (2012). Experimental Semiotics: An engine of discovery for understanding human communication. *Advances in Complex Systems*, *15*(3–4), 1150026. doi: 10.1142/S0219525911500263

Gordon, M. J. (2002). Investigating chain shifts and mergers. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.), *The Handbook of Language Variation and Change* (Vol. 11, pp. 244–266). Malden, MA/Oxford: Blackwell.

Hall-Lew, L. (2010). Improved representation of variance in measures of vowel merger. In *Proceedings of Meetings on Acoustics* (Vol. 9, p. 060002). Acoustical Society of America. doi: 10.1121/1.3460625

Haspelmath, M. (2006). Against markedness (and what to replace it with). *Journal of Linguistics*, *42*, 25–70. doi: 10.1017/S0022226705003683

Hay, J., Warren, P., & Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics*, *34*(4), 458–484. doi: 10.1016/j.wocn.2005.10.001

Hockett, C. F. (1960). The origin of speech. *Scientific American*, *203*, 88–96.

Honeybone, P., & Salmons, J. (2015). *The Oxford Handbook of Historical Phonology.* Oxford: Oxford University Press.

Hudson Kam, C. L., & Newport, E. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, *59*(1), 30–66. doi: 10.1016/j.cogpsych.2009.01.001

Jakobson, R., & Halle, M. (1956). *Fundamentals of Language.* The Hague: Mouton.

Kager, R. (1999). *Optimality Theory.* Cambridge: Cambridge University Press.

Keller, R. (2005). *On Language Change: The Invisible Hand in Language.* London: Routledge.

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. doi: 10.1073/pnas.0707835105

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, *28*, 108–114. doi: 10.1016/j.conb.2014.07.014

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102. doi: 10.1016/j.cognition.2015.03.016

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. doi: 10.18637/jss.v082.i13

Ladd, D. R. (2012). What *is* duality of patterning, anyway? *Language and Cognition*, *4*(4), 261–273. doi: 10.1515/langcog-2012-0015

Ladefoged, P., & Johnson, K. (2015). *A Course in Phonetics* (Seventh ed.). Stamford, CT: Cengage Learning.

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, *48*(4), 839–62. doi: 10.2307/411991

Lindblom, B. (2003). Patterns of phonetic contrast: Towards a unified explanatory framework. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)* (pp. 39–42).

Lindblom, B., & Engstrand, O. (1989). In what sense is speech quantal? *Journal of Phonetics*, *17*, 107–121. doi: 10.1016/S0095-4470(19)31516-5

Little, H., Eryılmaz, K., & de Boer, B. (2017). Signal dimensionality and the emergence of combinatorial structure. *Cognition*, *168*, 1–15. doi: 10.1016/j.cognition.2017.06.011

MacNeilage, P. F., & Davis, B. L. (2000). On the origin of internal structure of word forms. *Science*, *288*(5465), 527–531. doi: 10.1126/science.288.5465.527

Mantakas, M., Schwartz, J. L., & Escudier, P. (1986). Modèle de prédiction du 'deuxiéme formant effectif f_2'–application à l'étude de la labialité des voyelles avant du français. *Proceedings of the 15th journées d'étude sur la parole*, 157–161.

Micklos, A. (2016). Interaction for facilitating conventionalization: Negotiating the silent gesture communication of noun–verb pairs. In S. G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Fehér, & T. Verhoef (Eds.), *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11).*

Morin, O., Winters, J., Müller, T. F., Morisseau, T., Etter, C., & Greenhill, S. J. (2018). What smartphone apps may contribute to language evolution research. *Journal of Language Evolution*. doi: 10.1093/jole/lzy005

Mrayati, M., Carré, R., & Guérin, B. (1989). Distinctive regions and modes: A new theory of speech production. *Speech Communication*, *7*(3), 257–286. doi: 10.1016/0167-6393(88)90073-8

Nölle, J., Staib, M., Fusaroli, R., & Tylén, K. (2018). The emergence of systematicity:

How environmental and communicative factors shape a novel communication system. *Cognition*, *181*, 93–104. doi: 10.1016/j.cognition.2018.08.014

Nycz, J., & Hall-Lew, L. (2013). Best practices in measuring vowel merger. In *Proceedings of Meetings on Acoustics* (Vol. 20). doi: 10.1121/1.4894063

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Raviv, L., Meyer, A., & Lev-Ari, S. (2019). Compositional structure can emerge without generational transmission. *Cognition*, *182*, 151–164. doi: 10.1016/j.cognition.2018.09.010

Revelle, W. (2018). psych: Procedures for psychological, psychometric, and personality research [Computer software manual]. Evanston, Illinois. Retrieved from https://CRAN.R-project.org/package=psych

Roberts, G. (2017). The linguist's *Drosophila*: Experiments in language change. *Linguistics Vanguard*, *3*(1), 20160086. doi: 10.1515/lingvan-2016-0086

Roberts, G., & Fedzechkina, M. (2018). Social biases modulate the loss of redundant forms in the cultural evolution of language. *Cognition*, *171C*(1), 194–201. doi: 10.1016/j.cognition.2017.11.005

Roberts, G., & Galantucci, B. (2012). The emergence of duality of patterning: Insights from the laboratory. *Language and Cognition*, *4*(4), 297–318. doi: 10.1515/langcog-2012-0017

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. doi: 10.1016/j.cognition.2015.04.001

Sandler, W., Aronoff, M., Meir, I., & Padden, C. (2011). The gradual emergence of phonological form in a new language. *Natural Language and Linguistic Theory*, *29*, 502–543. doi: 10.1007/s11049-011-9128-2

Schwartz, J.-L., Abry, C., Boë, L.-J., Ménard, L., & Vallée, N. (2005). Asymmetries in vowel perception, in the context of the Dispersion–Focalisation Theory. *Speech Communication*, *45*(4), 425–434. doi: 10.1016/j.specom.2004.12.001

Schwartz, J.-L., Boë, L.-J., Vallée, N., & Abry, C. (1997). The dispersion-focalization theory of vowel systems. *Journal of phonetics*, *25*(3), 255–286. doi: 10.1006/jpho.1997.0043

Schwartz, J.-L., Boë, L.-J., Vallée, N., & Abry, C. (1997). Major trends in vowel system inventories. *Journal of Phonetics*, *25*, 233–253. doi: 10.1006/jpho.1997.0044

Sneller, B., & Roberts, G. (2018). Why some behaviors spread while others don't: A laboratory simulation of dialect contact. *Cognition*, *170C*, 298–311. doi: 10.1016/j.cognition.2017.10.014

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

Stevens, K. N., & Keyser, S. J. (2010). Quantal theory, enhancement and overlap. *Journal of Phonetics*, *38*(1), 10–19. doi: 10.1016/j.wocn.2008.10.004

Studdert-Kennedy, M., & Goldstein, L. (2003). Launching language: The gestural origin of discrete infinity. In M. H. Christiansen & S. Kirby (Eds.), *Language Evolution* (pp. 235–254). Oxford: Oxford University Press.

Thompson, D. W. (1942). *On Growth and Form.* Cambridge: Cambridge University Press.

Trendafilov, D., Lemmelä, S., & Murray-Smith, R. (2010). Negotiation models for mobile tactile interaction. In *International Workshop on Mobile Social Signal Processing.* Berlin/Heidelberg: Springer-Verlag. doi: 10.1007/978-3-642-54325-8_7

Vaux, B., & Samuels, B. (2015). Explaining vowel systems: Dispersion theory vs. natural selection. *The Linguistic Review*, *32*(3), 573–599. doi: https://doi.org/10.1515/tlr-2014-0028

Verhoef, T., Kirby, S., & de Boer, B. (2014). Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals. *Journal of Phonetics*, *43C*, 57–68. doi: 10.1016/j.wocn.2014.02.005

Virbel, M., Hansen, T., & Lobunets, O. (2011). Kivy – A framework for rapid creation of innovative user interfaces. In *Workshop-Proceedings der Tagung Mensch & Computer 2011. überMEDIEN|ÜBERmorgen.*

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer-Verlag. Retrieved from http://ggplot2.org

A Iconicity

Table 3 shows the full results, for all color components and referents, of the mixed effects models for iconicity (see Section 6.1). A different model was fit for each of the three color components, with the mean color value as the outcome variable, referent as predictor, condition as an interaction term, and dyad as a random factor. The results support the view that most referents were, at least in part, iconically motivated in form. However, there was no main effect of or interaction with condition.

Table 3: Results of mixed effects models for iconicity (df = 328).

	Mean red				**Mean green**				**Mean blue**			
Referent	Est.	SE	t	p	Est.	SE	t	p	Est.	SE	t	p
Bird	-0.01	0.1	-0.12	0.91	0.45	0.09	4.9	< 0.001	0.37	0.08	4.4	< 0.001
Butterfly	0.37	0.1	3.83	< 0.001	0.17	0.09	1.83	0.07	0.45	0.08	5.39	< 0.001
Deer	-0.04	0.1	-0.45	0.65	0.39	0.09	4.2	< 0.001	0.11	0.08	1.27	0.21
Dog	0.01	0.1	0.13	0.9	0.23	0.09	2.48	0.01	0.07	0.08	0.88	0.38
Dolphin	-0.17	0.1	-1.74	0.08	0.3	0.09	3.2	< 0.001	0.38	0.08	4.55	< 0.001
Duck	0.005	0.1	0.05	0.96	0.16	0.09	1.75	0.08	0.24	0.08	2.91	< 0.001
Elephant	-0.03	0.1	-0.35	0.73	0.11	0.09	1.17	0.24	0.05	0.08	0.55	0.58
Flamingo	0.16	0.1	1.59	0.11	0.06	0.1	0.59	0.56	0.3	0.09	3.35	< 0.001
Giraffe	0.13	0.1	1.36	0.18	0.2	0.09	2.08	0.04	0.07	0.08	0.78	0.43
Goat	-0.15	0.1	-1.48	0.14	0.28	0.09	3.01	< 0.001	0.08	0.08	0.91	0.36
Horse	0.03	0.1	0.3	0.77	0.03	0.09	0.29	0.77	0.09	0.08	1.09	0.28

B Calculating maximum pairwise distance

Calculating maximum possible pairwise distance analytically for a given inventory size is nontrivial; in fact the so-called maximum dispersion problem is well established in mathematics as being computationally difficult (Fernández, Kalcsics, & Nickel, 2013). We solved the problem for our purposes algorithmically for sets of up to 25 members (which was greater than any phoneme inventory in our data). One simple approach would be to successively add points to a space, such that each new point was maximally distant from currently existing points. For a set of size 25, the maximally dispersed set can be established by drawing three horizontal lines and three vertical lines that divide the whole space into 16 equally sized areas. (This can be thought of, alternatively, as dividing the space into successive halves.) The points where the lines meet are maximally dispersed with respect to each other. We did this for the dimensions of our production space and created a matrix of pairwise distances between all points. Then, beginning with a corner point (for a set of size one), we successively added points such that each new point maintained maximal mean pairwise distance with those already present.