# Signal Dimensionality and the Emergence of Combinatorial Structure

**Hannah Little[1,2] (hannah@ai.vub.ac.be), Kerem Eryılmaz[2] (kerem@ai.vub.ac.be) & Bart de Boer[2] (bart@arti.vub.ac.be)**

[1]Max Plank Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, the Netherlands

[2]Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium

In language, a small number of meaningless building blocks can be combined into an unlimited set of meaningful utterances. This is known as combinatorial structure. One hypothesis for the initial emergence of combinatorial structure in language is that recombining elements of signals solves the problem of overcrowding in a signal space. Another hypothesis is that iconicity may impede the emergence of combinatorial structure. However, how these two hypotheses relate to each other is not often discussed. In this paper, we explore how signal space dimensionality relates to both overcrowding in the signal space and iconicity. We use an artificial signalling experiment to test whether a signal space and a meaning space having similar topologies will generate an iconic system and whether, when the topologies differ, the emergence of combinatorially structured signals is facilitated. In our experiments, signals are created from participants' hand movements, which are measured using an infrared sensor. We found that participants take advantage of iconic signal-meaning mappings where possible. Further, we use trajectory predictability, measures of variance, and Hidden Markov Models to measure the use of structure within the signals produced and found that when topologies do not match, then there is more evidence of combinatorial structure. The results from these experiments are interpreted in the context of the differences between the emergence of combinatorial structure in different linguistic modalities (speech and sign).

**Keywords:** Signal Spaces, Artificial language experiments, Linguistic modalities, Linguistic Structure, Iconicity, Hidden Markov Models

## Introduction

Language is structured on at least two levels (Hockett, 1960). On one level, a small number of meaningless building blocks (phonemes, or parts of syllables for instance) are combined into an unlimited set of utterances (words and morphemes). This is known as *combinatorial structure*. On the other level, meaningful building blocks (words and morphemes) are combined into larger meaningful utterances (phrases and sentences). This is known as *compositional structure*. In this paper, we focus on *combinatorial structure*.

This paper investigates the emergence of structure on the combinatorial level. Specifically, we are interested in how the topology of a signalling space affects the emergence of combinatorial structure. We hypothesise that combinatorial structure will be facilitated when a meaning space has more dimensions (ways meanings can be differentiated) than the signal space has dimensions (ways signals can be differentiated). We are also interested in the emergence of iconicity. Iconicity is the property of language that allows meanings to be predicted from their signals. We posit that iconicity can also be facilitated by the topology of a signalling space, but when a meaning space and a signal space have similar numbers of dimensions, rather than differing ones. Taken together, these hypotheses will have different predictions for systems with different topologies. We posit that it is dimensionality that is at the root of why different signal structures may be facilitated by different linguistic modalities in the real world (speech and sign).

Previously, linguists have hypothesised that combinatorial structure is present in all human languages, spoken and signed (Hockett, 1960). Further, evidence suggests that at least in the hominid lineage, the ability to use combinatorial structure is a uniquely human trait (Scott-Phillips & Blythe, 2013). It therefore needs to be explained why human language has combinatorial structure. Hockett (1960) proposed that combinatorial structure emerges when the number of meanings, and therefore signals, grows, while the signal space stays the same. If all signals are unique (i.e. they do not overlap in the signal space), this means that the signal space becomes more and more crowded and that signals become more easily confused. Combining elements from a smaller set of essentially holistic signals into a larger set of longer signals makes it possible to increase the number of signals beyond what can be achieved by purely holistic signals. Others have hypothesised that combinatorial structure may be adopted as an efficient way to transmit signals when more iconic strategies are not available. Goldin-Meadow and McNeill (1999) propose that there is a relation between the emergence of combinatorial structure and the (in)ability for mimetic ($\approx$ iconic) signal-meaning mappings; spoken language needs to rely on combinatorial structure exactly because it cannot express meanings mimetically (iconically). Roberts, Lewandowski, and Galantucci (2015) argue that early in a language's emergence, if iconicity is available, this will be adopted over methods that are more efficient for transmission (such as combinatorial structure). This happens because iconicity is high in referential efficiency, which is more useful when languages are in their infancy, i.e. when linguistic conventions have not yet been firmly established in the language community.

An important source of evidence regarding the emergence of combinatorial structure comes from newly emerging sign languages, such as Al-Sayyid Bedouin Sign Language and Central Taurus Sign Language (Sandler, Aronoff, Meir, & Padden, 2011; Caselli, Ergin, Jackendoff, & Cohen-Goldberg, 2014). While these languages do combine words into sentences, the words they use do not appear to be constructed from combinations of a limited set of meaningless building blocks (e.g. handshapes). In other words: these languages do have compositional structure, but lack combinatorial structure (at least in the initial stages of their emergence). Conversely, it is not easy to imagine a spoken language without a level of combinatorial structure. Nothing similar has ever been reported for emerging spoken languages such as contact languages, pidgins and creoles. Taken together, these observations suggest that different linguistic modalities cause differences in how structure emerges. Here we ask whether this is due to the availability of more iconicity in signed languages, or a constraint in the amount of distinctions possible in spoken languages.

## Signal-space crowding and the emergence of combinatorial structure

Mathematical models (Nowak, Krakauer, & Dress, 1999) and computational models (Zuidema & de Boer, 2009) show that combinatorial signals can indeed theoretically emerge from holistic signals as a result of overcrowding in the signal space. However, in reality, the process of transition from holistic to combinatorial signals involves more factors. The evidence from emerging sign languages mentioned above shows that apparently fully functional languages can get by without combinatorial structure. These emerging languages slowly transition from a state without combinatorial structure to a state with combinatorial structure, without a marked increase in vocabulary size (Sandler et al., 2011). Apparently, the size and flexibility of the sign modality allows for a fully holistic language (on the word level) in an initial stage.

Backing up the naturalistic results, and in contrast with the models, experimental investigations have failed to show a strong correlation between the crowdedness of the signal space and the emergence of combinatorial structure. Verhoef, Kirby, and de Boer (2014) investigated the emergence of structure in sets of signals that were produced with slide whistles. Participants learnt a set of 12 whistled signals, and after a short period of training, their reproductions were recorded and used as learning input for the next "generation" of learners. This process of transmission from generation to generation was modelled in an iterated learning chain of 10 generations (Kirby, Cornish, & Smith, 2008). They found that even in this small set of signals, combinatorial structure emerged rapidly and in a much more systematic way than through gradual shifts as predicted by Nowak et al. (1999) and Zuidema and de Boer (2009). This indicates that processes of reanalysis and generalisation of structure play a more important role than just crowding of the signal space.

Roberts and Galantucci (2012) also investigated whether crowding in the signal space affected the emergence of combinatorial structure. Participants developed a set of signals to communicate about different animal silhouettes. The instrument used to generate graphical signals (designed by Galantucci, 2005) prevented them from either drawing the silhouettes, writing the name of the animals, or using other pre-existing symbols. They found that there was no strong relation between the number of animals communicated by participants and the level of structure found in signals.

Little and de Boer (2014) adapted Verhoef et al's (2014) slide whistle experiment to investigate how the size of the signal space would affect the emergence of structure. By limiting the movement of the slider of the slide whistle with a stopper, the possible signals were restricted to a third of the original pitch range. There was no significant difference in the emergence of structure between the reduced condition and the original condition, indicating that there was no strong effect of reducing the available signal space on the emergence of combinatorial structure. However, although the stopper prevented a certain portion of the pitch range from being used, it did not affect participants' ability to replicate essential features of the trajectories that could be produced without a stopper (for example, a rising pitch repeated). With the specific example of slide whistle signals, it is not the size of the signal space that would cause overcrowding, but the way in which signals in the space can be modified and varied. This idea is at the core of the present work and will be discussed more thoroughly below.

The current experimental evidence, then, seems to suggest that crowding in the signal space does not play such a primary role in the emergence of structure as predicted by Hockett. However, it is clear that the nature of the signal space must influence the emergence of combinatorial structure, otherwise, we could not explain that the sign languages can exist (at least briefly) without combinatorial structure, whereas spoken languages apparently cannot. One reason for this difference between modalities could be the extent to which a given signalling medium allows for the use of iconicity.

## Iconicity and Combinatorial Structure

Hockett (1960) proposed that an arbitrary mapping between signal and meaning is a design feature of language. However, it is now well-accepted that there is a non-trivial amount of iconicity in human language. In spoken language, the most salient example is true onomatopoeia, the property that a word sounds like what it depicts (e. g. cuckoo, peewit, chiffchaff and certain other bird names), though this is quite rare. A more common form of iconicity is sound symbolism, which has now been demonstrated to be much more widespread than previously thought (Blasi, Wichmann, Hammarström, Stadler, & Christiansen, 2016). In sound symbolism, there is a less direct relation between the signal of a word and its meaning than in onomatopoeia. One example is that of the relation between the size of an object that a word indicates and the second formant of the vowel(s) it contains. Vowels with a high second formant tend to indicate smallness, as in words like "teeny" (Blasi et al., 2016). Another very different example is that words that start with sn- often have something to do with the nose: sneeze, sniff, snot, snout etc. (possibly because "sn" is onomatopoeic for the sound one makes when one has a cold). Here sn- almost functions like a morpheme, but its meaning is not sufficiently well-defined to be a true morpheme, and there are many words starting with sn that have nothing to do with the nose. In sign languages, there is a lot of visual iconic structure. For instance, the sign for tree in British Sign Language has the arm representing the trunk, with the fingers pointing upwards and splayed to represent the branches of the tree. Although it is hard to quantify precisely, iconic structure is more prevalent in sign language than in spoken language. This assumption is supported by experimental evidence demonstrating that it is more difficult to be iconic using vocalisations than it is with gestures (Fay, Lister, Ellison, & Goldin-Meadow, 2014). Further, sign languages have more signal dimensions than spoken languages (Crasborn, Hulst, & Kooij, 2002). More signal space dimensions allow for more mappings to be made between the signal space and the highly complex meaning space we communicate about in real life, especially when those meanings are visual or spatial in nature.

In the introduction we mentioned the hypothesis of Goldin-Meadow and McNeill (1999) and Roberts et al. (2015); that iconicity suppresses the emergence of combinatorial structure. Roberts and Galantucci (2012) explore how this mechanism could work. They hypothesise that as signs become conventionalised, iconicity may become dormant, i.e. language users are no longer aware of it. Once iconicity has been lost (or become dormant) through a process of conventionalistion, this opens up the possibility of re-analysing regularities in signs as meaningless building blocks that then become standardised across signs. Iconic signs are robust to variation, as their meaning can be compensated for with knowledge of the world. This is not possible when signs or building blocks become arbitrary, and so a pressure for all speakers to adhere to the same standard takes over. These hypotheses suggest that the ability to use iconicity interacts with the emergence of combinatorial (and compositional) structure.

Evidence for the connection between iconicity and combinatorial structure comes from several recent experimental studies. Roberts and Galantucci (2012) found in their animal silhouette experiment that more iconic signals tend to be less combinatorial. Further, Roberts et al. (2015) conducted a study where the meanings could either be easily represented iconically or not, with the results indicating the emergence of combinatorial structure in non-iconic signals, but not in those that retained their iconicity. Similarly, Verhoef, Kirby, and Boer (2015) showed that structure emerged differently in a situation where participants could make use of possibly iconic signal-meaning mappings than in a situation where they could not. The experiment used the same setup as the one described above (Verhoef et al., 2014), except that the whistles were associated with meanings. In one condition, signals were paired with the same meaning they were produced for when passed to the next generation for learning. This meant that iconicity in signals could persist in transmission. In the other condition, a random meaning was associated with each unique signal presented to the listener, so that producer and listener did not have the same meaning for a given signal. The former condition allowed for transmission of iconic signal-meaning mappings, while the latter condition did not. Verhoef et al. (2015) found that structure emerged faster in the condition where signal-meaning mappings were not preserved, i.e. where iconicity was not possible.

In the experiments above, iconicity is either possible or not. However, the difference in iconic ability between spoken and signed language is one of degree rather than a parameter that is "on" or "off". In the experiments in the current paper, we are interested in how more nuanced manipulations of available signal-meaning mappings can promote the emergence of combinatorial structure.

## The Current Study

### Iconicity in the current study

In this paper, we investigate whether the observed differences in the emergence of structure are dependent on the degree of iconicity a particular signal space affords. Iconicity can take various forms, as we have already made clear. However, we need to formalise notions of different types of iconicity in order to inform our experimental design and results. We define two forms of iconicity: relative and absolute iconicity (Monaghan, Shillcock, Christiansen, & Kirby, 2014). For relative iconicity, there is what mathematicians call a homeomorphism between the meaning space and the signal space (i.e. there is an invertible mapping in which neighbouring points in the meaning space stay neighbouring points in the signal space). The consequence of such a mapping is that if one knows enough signal-meaning mappings (at least the number of dimensions +1), then meanings corresponding to unseen signals and signals corresponding to unseen meanings can be guessed. In order for this mapping to work, points along the dimensions of the meaning and signal spaces must be ordered in some way. Meaning and signal spaces with categorical dimensions (e.g. biological sex) do not allow for such generalisable relative iconicity. Indeed, previously, we conducted an experiment using continuous signals to refer to meanings with categorical dimensions (Little, Eryılmaz, & de Boer, 2015). Using the same methodology as the current paper (see Methods section below), we compared what happens when a continuous signal space is used to describe a continuous meaning space verses a discrete meaning space. We found that the discrete condition created signals with more movement and structure when relative iconicity was more difficult. This suggests that structure may emerge due to transparent mappings not being available, which fits with the findings from the experiments mentioned above (Roberts & Galantucci, 2012; Roberts et al., 2015; Verhoef et al., 2015).

For absolute iconicity, one only needs to see one signal in order to see an iconic relation. To achieve this, the dimensions that correspond through the homeomorphism must also correspond to a feature in the real world. For example, this is the case in the absolute iconic mapping between the second formant of vowels [i], [o], [u] and size, where the second formant (a frequency) maps to the pitch that an object would make if tapped. It should be noted that these dimensions do not have to be linear and continuous. They can be spatial (as in directions) or discrete/categorical (as in presence and absence of a property). In addition, similarity is a very broad notion in practice; it often takes the form of an associative link between a property (e.g. size) and a selected feature that corresponds to that property (e.g. frequency when tapped). Depending on the number of dimensions that are related to the same feature in the real world, the indirectness of these links, and the total number of dimensions that are mapped through the homeomorphism, there is a continuum between absolute iconicity, relative iconicity and no iconicity at all.

## Topology in the current study

In our experiments, the notion of topology allows us to operationalise the way signal and meaning spaces map onto each other. When a meaning space has the same number of dimensions (or fewer) as the signal space, an iconic mapping is possible. When the number of dimensions of the signal space is lower than that of the meaning space, completely iconic mappings are no longer possible.

Zuidema and Westermann (2003) were the first to look at signal and meaning spaces with identical topologies. They looked at meanings and signals from a bounded linear space. Using a computer simulation, they found that the most robust signal-meaning mapping was a topology-preserving iconic mapping: one in which signals that were close together corresponded to meanings that were close together. In this way, small errors in production and perception only disrupted communication minimally. In a follow-up study, de Boer and Verhoef (2012) found that, while this works when the topologies of the signal and meaning space match, when the meaning space has more dimensions than the signal space, mappings emerge that show structure. Here, we propose that de Boer & Verhoef's (2012) model can inform us about the emergence of structure in signed and spoken language: the signal space of signed languages (in comparison to the signal space of spoken language) is closer in topology to the (often visual and spatial) meaning space that humans tend to talk about. The more overlap there is between topologies, the easier it is to find signal-meaning mappings where a small change in signal corresponds to a small change in meaning. Moreover, when the topologies map, it is possible to have productive iconic signal sets where new signals are predictable from existing ones (for instance, higher pitches corresponding to smaller objects). In order to develop these ideas further, it is first necessary to experimentally investigate whether the effects predicted by de Boer and Verhoef (2012) hold for human behaviour.

In our experiments, we manipulate the number of dimensions in our signal and meaning spaces to investigate the properties of the signalling systems that participants create. The number of dimensions (the dimensionality) of the meaning space is manipulated by varying images in size, shade and/or colour. The number of dimensions in the signal space is controlled by using an artificial signalling apparatus (built using a *Leap Motion* infra-red hand position sensor) that produces tones that can differ in intensity and/or pitch depending on hand position. This allows us to have different combinations of signal and meaning space dimensionality, and therefore different mappings between the topologies of these spaces.

One important implication to manipulating the topology of our signal space is that dimensionality is not only tied to the iconicity possible (as outlined above), but it also affects the size of a signal space. The more dimensions a signal space has, the more distinctions can be made between signals in that space. This means that the overcrowding of signal space hypothesis and the iconicity hypothesis cannot be teased apart by the experimental work in this paper directly. They may also be more interrelated in real world languages than is indicated in previous work.

## Experiments

Our experiments aim to explore the effects that signal space topology has on the emergence of structure. Specifically, following the themes of de Boer and Verhoef (2012), we aim to find out how differences in the dimensionality of both the signal space and the meaning space will affect the structure in signals used. Following the findings of de Boer and Verhoef (2012), our hypothesis is that when the dimensionality of the signal space is lower than that of the meaning space, then combinatorial structure will be adopted. We also expect that when there is matching dimensionality in signal and meaning spaces, then participants will adopt iconic strategies.

Experiment 1 compares signal spaces which are either 1 dimensional (pitch or volume) or two-dimensional (both pitch and volume). These signals were used to label meanings that either differed in only one dimension (size) or two dimensions (both size and shade of grey). However, we found that participants used duration as a signal dimension, meaning that the number of signal dimensions did not correspond to the intended number in the experimental design. To fix this, in Experiment 2, signals only differed in pitch (and duration) and the meaning space grew to 3 dimensions to ensure we could observe the effects of meaning dimensions outnumbering signal space dimensions.

## Experiment 1

Experiment 1 consisted of signal creation tasks and signal recognition tasks. In contrast to previous experimental work, these signals were not used for communication between participants, or iterated learning. Instead, participants created and then recognised their own signals.

## Methods

**Participants.** Participants were recruited at the Vrije Universiteit Brussel (VUB) in Belgium. 25 participants took part in the experiment; 10 male and 15 female. Participants had an average age of 24 (*SD* = 4.6). No participants reported any knowledge of sign languages. We also asked participants to self-report their musical proficiency (on a scale of 1-5). This information was recorded as recognition of pitch-track signals might be dependent on participants' musical abilities, so we needed to identify and control for this potential effect in our results.

**The signal space.** Our experiment used a continuous signal space created using a *Leap Motion* device: an infrared sensor designed to detect hand position and motion (for extensive details about the *Leap Motion* paradigm, see Eryılmaz & Little, 2016). Participants created auditory signals using their hand positions within the space above the sensor. The *Leap Motion* was used to generate continuous, auditory signals that were not speech-like. In this way, we could see how structure emerged in our signals in a way that is analogous to speech, without having pre-existing linguistic knowledge interfere with participants' behaviour.

We could manipulate the dimensionality of this signal space, so signal generation depended on moving the hand within a horizontal dimension (x), vertical dimension (y) or both (Figure 1). Signals were generated that either differed in pitch (on the x-axis), volume (on the y-axis), or both. Participants were told explicitly which signal dimension(s) they were manipulating. When a signal could be altered along two perceptual dimensions (i.e. pitch and volume), participants achieved this by moving one hand within a two-dimensional space, i.e. moving a hand up or down would affect the volume, while a hand moving left or right would manipulate the pitch. Participants could hear the signals they were producing. Participants were given clear instructions on how to use the sensor and had time to get used to the mapping between their hand position and sound.

Both the pitch and volume scales used were non-linear. Though our paradigm allows for any mapping between the hand position and the acoustic signal, participant feedback in pilots indicated that people could more intuitively manipulate non-linear scales. However, the output data has variables for both absolute hand

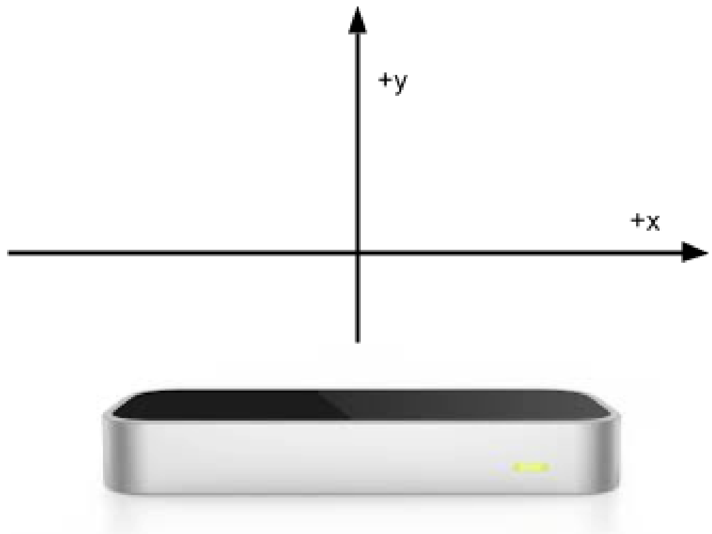


*Figure 1*. The signal dimensions available using the *Leap Motion*. In phases with a one-dimensional signal space, only either the x- or y-axis was available.

position within signal trajectories (represented as coordinates), and transformed pitch and volume values so that we could explore whether participants were relying more on hand position or the acoustic signal.

Recording was interrupted when participants' hands were not detectable, meaning that there were no gaps in any of the recorded signals, even if participants tried to produce them. This was done to stop participants creating gaps to separate structural elements in the signals, as this is not something typically used to separate combinatorial elements in speech or sign. The data does not show much (if any) evidence that participants tried to include gaps in the experimental rounds, which would be evident from sudden changes in pitch in the signal.

**The meaning space.** The meaning space consisted of a set of squares that differed along continuous dimensions. In phases where the meaning space only differed on one dimension, five black squares differed only in size. In phases where the meaning space differed on two dimensions, nine squares differed in both size and in different shades of grey (Figure 2). Participants had to create distinct signals for each square.

**Procedure.** Participants were given instructions on how to generate signals using the *Leap Motion*. They were given time to practice using the *Leap Motion* while the instruction screen was showing. Participants had control of when to start the experiment, and so could practice for as long as they wanted. They were instructed to sit back in the chair during the experiment, so that their upper body did not interfere with the *Leap Motion*. Participants were also told that they would have to recognise the signals they produced, so they knew they had to make signals distinct from one another.

There were three phases of the experiment: each phase consisted of a practice round and an experimental round. There was no difference between practice rounds and experimental rounds, but only the data from the experimental round was used in the analysis. Each practice and experimental round consisted of a signal creation task and a signal recognition task.

**Signal Creation Task.** At the beginning of each signal creation task, participants saw the entire meaning space. They then were presented with squares in a random order, one by one, and pressed an on-screen button to begin and finish recording their signals. They had the opportunity to play back the signal they had just created, and rerecord the signal if they were not happy. Participants created signals for all possible squares in a phase.

**Signal Recognition task.** After each signal creation task, participants completed a signal recognition task. All signals they had created were presented to them in random order one after the other. For each signal, they were asked to identify its referent from an array of three randomly selected meanings (from the repertoire of possible meanings - i.e. squares of different colours and shades of grey - within the current phase) plus the correct referent, so four meanings in total. They were given immediate feedback about whether they were correct, and if not, what the correct meaning had been. This task worked as a proxy for the pressure to communicate each meaning unambiguously (expressivity), as participants knew that they had to produce signals that they could then connect back to the meaning in this task, thus preventing them from producing random signals, or just the same signal over and over again. Their performance in this task was recorded.

When participants were incorrect, we measured the distance between the meaning they selected and the correct meaning. The distance was calculated as the sum of differences along each dimension using a measure similar to Hamming distance. Let $m_{ij}$ define a meaning with size $i$ and shade $j$ in a meaning space where $0 < i < I$ and $0 < j < J$ . The distance between two meanings $m_{ij}$ and $m_{i'j'}$ is then the following:

$$D(m_{ij}, m_{i'j'}) = \left|i - i'\right| + \left|j - j'\right| \tag{1}$$

For example, if the correct square has values 3 and 3 for size and shade respectively, and the chosen square had vales 1 and 2 for size and shade respectively, the distance between these two squares would be 3. Correct answers have a distance of 0.

**Phase 1:1.** All participants started with phase 1:1. In this phase, the meaning space consisted of five black squares, each of different sizes (one meaning dimension). In this phase, the signal space also had only one dimension, which was either pitch or volume. Which signal dimension the participants started with was assigned at random. This phase was a matching phase, as there was a one to one mapping possible between the meaning space and signal space (Figure 2).

**Phase 1:2.** In phase 1:2, participants created signals for a two-dimensional meaning space with the squares differing in size and shade. The signal space had only one dimension. Participants used the same one-dimensional signal space that they used in phase 1:1, so if they started the experiment only using pitch, they only used pitch in this phase. This was the mismatch phase, as there were more meaning dimensions than signal dimensions (Figure 2).

**Phase 2:2.** In phase 2:2, participants described the two-dimensional meaning space (differing in size and shade), but with a two-dimensional signal space, where the signals differed in both pitch and volume along the x and y dimensions respectively (Fig. 2). This phase was a matching phase also, as there was a one to one mapping available between signal and meaning spaces.

**Counterbalancing.** Participants completed the phases in order 1:1, 1:2, 2:2 (where mismatch phase interrupts matching phases) or 1:1, 2:2, 1:2 (where matching phases are consecutive). Order was counterbalanced because participants' behaviour may depend on what they have previously done in the experiment. If people must solve the dimensionality mismatch before being presented with the two-dimensional signal space, then they may continue using an already established strategy that only uses only one dimension, rather than change their strategy to take advantage of both dimensions.

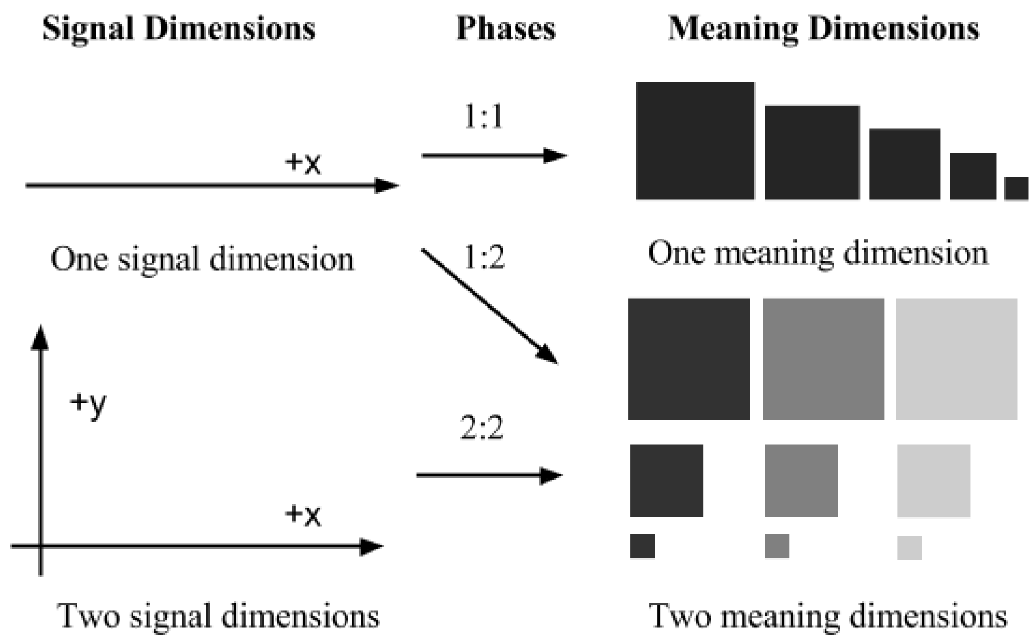


*Figure 2*. The phases used in the experiment. Phase 1:2 is the mismatch phase.

**Post-experimental questionnaire.** We administered a questionnaire with each participant after they had completed the experiment. This questionnaire asked about the ease of the experiment, as well as about the strategies that the participant adopted during each phase of the experiment. The questionnaire asked explicitly whether they had a strategy and, if so, how the participant encoded each meaning dimension into their signal.

## Results

### Signal Creation Task

The data collected from the signal creation task consisted of coordinate values designating hand position at every time frame recorded, which is what the following statistics are based on. There were approximately 110 time frames per second. Signals were on average 3.36 seconds long. We first looked at the mean of the coordinate values for each trajectory, and the duration of each signal. These simple measures give a good starting point to assess whether participants were encoding the meaning space directly with the signal space. If size or shade was directly encoded by pitch, volume or duration trough relative iconicity, then this should be detectable in the mean coordinates or duration of the trajectories.

The first dimension a participant used was collapsed into one outcome variable in our analysis, regardless of whether it was pitch or volume. All coordinates for signals using either pitch or volume were normalised to have the same range. We also controlled for whether these coordinates were pitch or volume in the mixed linear models below as a fixed effect, and also ran a separate analysis that showed that participants performed just as well in the task when starting with either pitch or volume (reported in the signal recognition results below). As explained above, meaning dimensions were coded to reflect the continuous way in which they differed, i.e. the smallest square was coded as having the value of 1 for size, and the biggest square a value of 5, with the lightest grey square given a value of 1 for shade, and the darkest had a value of 3. Using these values, we could predict duration and mean coordinates from size and shade.

We ran a mixed linear model with size and shade as predictors, duration and mean coordinate value as outcomes. Participant number was included as a random effect, and whether their starting dimension was pitch or volume as a fixed effect. P-values were obtained by comparing with null models that did not include the variable of interest. In the first phase, duration was predicted by the size of the squares ($\chi^2(1) = 18.5, p < 0.001$), but the mean coordinate value was not. In the other 2 phases, the mean coordinate of signals on the first dimension that a participant saw in phase 1:1 (either pitch or volume) was predicted most strongly by shade. A mixed linear model, controlling for the same effects as above, showed this interaction to be significant ($\chi^2(1) = 341.4$, $p < 0.001$). The duration of the signal was predicted most strongly by the size of the square, with each step of size increasing the signal by 75.296 frames$\pm$7(std errors) (approx 0.7 seconds). The mixed linear model for this effect, controlling for the same fixed and random effects, was also significant ($\chi^2(1) = 103.14$, $p < 0.001$). These effects demonstrate a propensity for encoding the meaning space with the signal space using relative iconicity. Size and duration are easy to map on to one another, and it makes sense that participants are more likely to encode the remaining meaning dimension (shade) with the signal dimension they were first exposed to. Figure 3 shows the output of one participant who mapped the signal space onto the meaning space in a very straightforward one to one mapping, with size encoded with duration and shade encoded with volume. This is an example of a topology-preserving mapping (a homeomorphism).

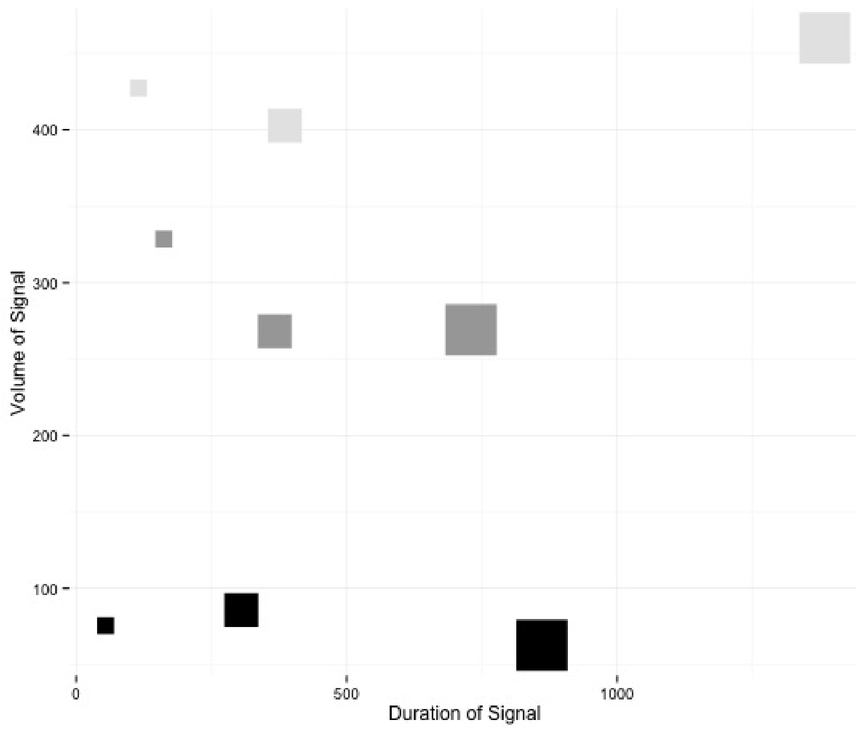


*Figure 3*. The mean trajectory coordinates (in mm) along the axis manipulating volume (where lower values refer to louder sounds) plotted against duration (number of data frames, roughly 1/110 of a second). Size and shade are represented by the size and shade of the squares in the graph. Within the phase with the two-dimensional meaning space with a two-dimensional signal space, this participant used signal duration to encode size, and signal volume to encode shade.

We also looked at standard deviation in signals to give us a good idea of the amount of movement in a signal. Signal trajectories produced in the phase where there was a mismatch (1:2) had higher standard deviations ($M$ = 48.2mm) than signals produced in phases where the signal and meaning spaces matched in dimensionality ($M$ = 33.4mm), indicating more movement in mismatch phases. Using a linear mixed effects analysis with standard deviation as the outcome variable and whether phases were matching or mismatching as the predictor, and controlling for participant number as a random effect, and whether they started with pitch or volume as a fixed effect, we found a significant effect ($\chi^2(1) = 4.5$, $p < 0.05$).

**Predictability of signal trajectories**

We also quantified signal structure by measuring the predictability of signal trajectories given other signals in a participantâĂŹs repertoire. If each signal trajectory in a participant's repertoire is predictable from the other signals, this gives an indication of systematic and consistent strategies being used within the repertoire.

We created a measure for predictability of each signal trajectory, derived from a participant's entire repertoire. The procedure is as follows:

1. Use the k-means algorithm to compute a set of clusters $S$ of hand coordinates using the whole repertoire, which reduce the continuous-valued trajectories to discrete ones ($k = 150$).

2. Calculate the bigram probability distribution $P$ for each symbol $x_i \in S$.

3. Use the bigram probabilities to calculate the negative log probability of each trajectory using Equation 2 below.

The choice of $k$ was set quite high at 150 to ensure the quantisation was sufficiently fine-grained. This ensured that the high variation in our data set is well-represented in the prediction scores to avoid overestimating similarity. In the literature, such high values for this parameter are used for modelling high-dimensional speech data, which we used as an upper bound (e.g. Räsänen, Laine, & Altosaar, 2009).

Letting $S$ be the set of 150 clusters obtained in step 1, and $T$ be a trajectory that consists of $m$ symbols $x_0, x_1, x_2, \ldots, x_m$ where $x_i \in S$, the formal description of step 3 is the following:

$$\mathrm{P(T)} = -\mathrm{log}P(x_0) - \sum_{a=1}^{m} logP(x_a|x_{a-1}) \tag{2}$$

With the predictability value for each trajectory, we used a linear mixed effects model to compare the predictability of signals in the matching and mismatching phases. Controlling for duration and participant number as random effects, and size and shade of square as fixed effects, we found that whether signals were produced in matched or mismatched phases predicted how predictable a trajectory was ($\chi^2(1) = 3.9$, $p < 0.05$). Signals produced in the matching phases had higher predictability.

### Signal Recognition Task

Overall, participants were good at recognising their own signals, identifying a mean of 66% of signals correctly, where 25% was expected if participants performed at chance level. Using a linear regression model, we found that participants improved by around 10% with each phase of the experiment ($F(1,76) = 9.96$, $p < 0.01$).

There was no significant difference between the recognition rates of participants who started with either volume or pitch ($t(21.9) = -0.46$, $p = 0.65$), suggesting that there was no difference in difficulty between the signal dimensions. We also used a linear regression model to test if musical proficiency predicted performance in the signal recognition task, and found that it did not ($F(1,23) = 0.03, p = 0.86$).

If signals rely on relative iconicity, then similar signals will be used for similar meanings, causing more potential confusion between signals for similar squares. This confusability may cause participants to be worse at the signal recognition task when relative iconicity is more prevalent. We tested whether participants were indeed worse at the recognition task in the condition where we predicted relative iconicity (in the matching phases). In line with this hypothesis, we found that participants were worse at recognising their signals within matching phases (1:1, 2:2) (*M* = 61.3% correct, *SD* 24%), than in mismatching phases (1:2) (*M* = 69.6%, *SD*= 21%). However, this result was not significant ($t(53.3) = -1.5$, $p = 0.13$), and may be an artefact of the experiment getting more difficult as it progressed.

We also calculated the distances between incorrect answers and target answers, as discussed in our methods (Signal Recognition Task section). To compare these values to a baseline, we also calculated the distance between the target answer and a randomly chosen incorrect answer. Comparing the actual data with the random data using a mixed effect linear model, and controlling for participant number as a random effect, and stimulus number as a fixed effect, we found that with incorrect choices produced in the matching phases (1:1, 2:2), participants were closer to the correct square (*M* = 2.6 steps away, *SD*= 1.4) than if they had chosen at random ($M$ = 3 steps away, $SD$= 1.7) ($\chi^2(1) = 5.5$, $p = 0.02$). However, in the mismatching phase (1:2) there was no difference between actual incorrect choices and random incorrect choices (both around 3.6 steps away, $\chi^2(1) = 0.01$, $p = 0.9$). Further, we found that the distance from the correct answer was much higher in the mismatching phases ($M$ = 3.6 steps away, $SD$ = 1.5), than in the matching phases ($M$ = 2.6 steps away, $SD$ = 1.4), indicating that participants were relying more on relative iconicity in the matching phases, because their mistakes were predicable, assuming a transparent mapping between the signal space and the meaning space. We tested this using a mixed effect linear model, and controlling for the same variables found the effect was significant ($\chi^2(1) = 5.3$, $p < 0.05$).

**Post-experimental questionnaire**

Nearly all participants reported strategies and they were mostly the same strategies. These strategies included using pitch, volume or duration directly to encode size or shade. For example, many participants used high pitches or short durations for small squares and low pitches or long durations for big squares. Participants also reported that involved different movement types, frequencies and speeds.

As we predicted in the section on counterbalancing, participants who saw phase 1:2 before phase 2:2, were more likely to use the same signal strategy throughout, than to change the strategy to take advantage of both dimensions. 84% ($SD$ = 37%) of strategies used for a particular meaning dimension were consistent throughout phases 1:2 and 2:2 by participants who saw 1:2 first. Only 54% ($SD$ = 50%) of strategies by those who saw 2:2 first were consistent. Consistency rates between different phase orders were significantly different ($\chi^2(1) = 8.7$, $p < 0.01$).

Whether a participant self-reported as having a strategy or not influenced their performance in the signal recognition task. Participants were significantly more likely to perform better at recognising their own signals in a given phase, if they reported having a strategy ($M = 70\%$ correct, $SD$ = 20%), than if they didn't ($M = 40\%$ correct, $SD$ = 16%) ($t(26.6) = -6$, $p > 0.001$).

**Hidden Markov Models**

**Models.** While our predictability values outlined in the previous sections are useful to characterise internal similarities in a repertoire, the clustering algorithm they are based on ignores temporal dependencies. To infer the structure of the signal repertoires including the temporal dependencies, we used Hidden Markov Models, or HMMs. An HMM consists of a set of *states* of which only one can be *active* at a time. The active state produces observable *emissions* (such as short stretches of time) drawn from a state-specific distribution, and the next active state depends only on the currently active state. In the models we derive from our experimental data, states are analogous to phonemes (or similar to building blocks), and the emission distributions to determine how they are realised phonetically. By training HMMs on the signal repertoires, we can estimate the most likely vocabulary of states across a repertoire, i.e. the most likely "phonological" alphabet. Note that this model does not explicitly include meanings, since our purpose is to model the structure of the signal repertoire.

HMMs are very common in natural language processing applications, such as part-of-speech tagging and speech recognition (Baker et al., 2009). A common use for HMMs in the field is modelling phonemes, where typically three states represent three phoneme positions, and their emissions are very short segments of speech making up the observed signal (see *Figure 4*).

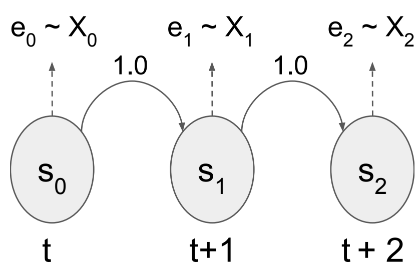


*Figure 4*. A simple, three state, left-to-right HMM emitting the observation sequence $e_0e_1e_2$ through the state sequence $s_0s_1s_2$. Each observation $e_i$ is a random sample from the emitting state's emission distribution $X_i$ where $i \in \{0,1,2\}$. Transitions are annotated with their probabilities. Note how the only non-deterministic part of the system is the emissions in this type of HMM.

HMMs are typically used with a fixed transition matrix and a fixed number of states. Each phoneme is modelled as a "left-to-right" HMM. These models have exactly one possible starting state, and all transitions are deterministic. Further, applications typically assume the number of states is already known and only the emission distribution for each state needs to be estimated. While this is useful for modelling a signal whose structure is familiar (such as human speech), it is not a very useful method of discovering and/or characterising structure in signals where the properties of the signalling system are unknown. Most of the structural variation available is ruled out by the fixed architecture of the HMM. Furthermore, contrary to common practice, we are interested in modelling the properties of the whole signal repertoire rather than individual signals.

Since we use HMM as a model of the speaker, the estimated properties of the model should be able to predict the participant's performance, such as their score in the recognition task for that phase. In particular, we are interested in whether the number of states in the HMM can predict the recognition score of a participant. Since the states are analogues for the phonemic inventory, we predict participants with bigger inventories will have worse recall. Such predictive power would indicate the model successfully captures aspects of participant behaviour during the experiments.

We propose that fewer building blocks across a repertoire indicates combinatorial strategies in comparison to strategies of relative iconicity. The efficiency that combinatorial structure brings is due to its capability to encode multiple meanings with combinations of a limited number of fundamental building blocks (or states in the HMMs). We expect combinatorial strategies (represented by a smaller numbers of states) to be more efficient in communicating meanings, because they overcome the problem of crowding in the signal space resulting in less confusion between signals. On the other hand, a system with relative iconicity, which would have to maintain a systematic relationship between the meanings and forms, would result in many states within a crowded system. With a combinatorial system, encoding a newly encountered meaning dimension does not require the invention of a new signal dimension to provide a range of signals to encode variations on the meaning dimension, which is what would happen with relative iconicity. We predict that the signals from phases where the number of meaning dimensions is greater than the number of signal dimensions will have combinatorial structure, and this will manifest itself in HMMs trained on those signals having *fewer* states than signals from matching phases.

We calculate the structure as well as the transitions of HMMs, with only an upper boundary on the number of states and no constraints on transitions. We use HMMs with continuous multivariate (Gaussian) emissions and the standard Baum-Welch algorithm for unsupervised training. We trained a separate HMM on the set of signals generated by each participant at each phase of the experiment. This way, we ensured that all signals that went into training a particular HMM had been created to label the same meaning space.

Because the mapping between hand position and the tones generated is non-linear, it makes a difference to the HMM which representation we use to train it. Which one works best depends on how participants memorise signals. There is no way of knowing *a priori* whether the participants will memorise (and when playing as the hearer, reverse-engineer) the hand movements themselves, or the tones produced by these movements. So, in addition to the raw data that assumes the states emit hand coordinates, we trained the models on two transformed data sets that assume the emissions are tonal amplitude and frequency values. These two additional sets varied in their frequency units, one using the Mel scale and the other Hertz. The full training procedure used for each projection is presented in *Algorithm 1* in *Appendix A*.

A series of linear mixed effects regressions were run to see what aspects of the HMMs are most useful in predicting the signal recognition scores. The dependent variable and covariates we have considered are the number of states of the model, while the predictors were phase, phase presentation order, and whether the phase is matching or mismatching. The random effects were whether volume or pitch was the first signal dimension introduced, and the participant number. Likelihood ratio tests were used to justify every additional component to the regression equation, corrected for the number of comparisons. The details of the regression and estimated coefficients are in *Appendix B*. Phases are coded as $p \in \{1:1, 1:2, 2:2\}$, independent of their presentation order (see *Counterbalancing* in the *Methods* section for explanation about order of the phases). Order of presentation is taken into account in the analysis, and is coded as "consecutive" (when the matching phases appear one after the other) or "interrupted" (when the mismatching phase appears between the matching phases). The matching phases are $p \in \{1:1, 2:2\}$, and the mismatching phase is $1:2$.

**Experiment 1 HMM Results and Discussion.** The interaction of number of states, phase order and mismatch was the best predictor for participant score in each phase ($R^2 = 0.616$). The signal representation most successful in predicting the recognition score was the Mel frequency and the amplitude in linear scale.

All results reported here come from HMMs trained on trajectories represented in Mel.

Some combinations of the interacting components were logically excluded; for instance, the 1:1 phase can only take place in the first position, so there is no coefficient for the interaction between the 1:1 phase and phase orders other than 1. See *Figure 5* for the regression coefficients.

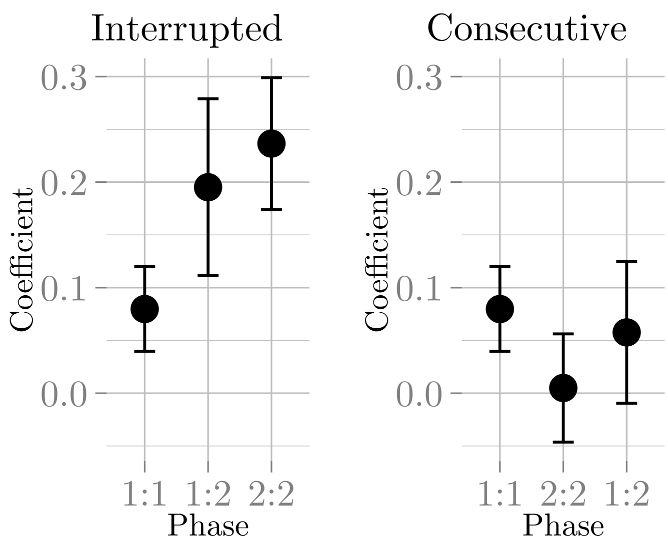


*Figure 5*. Fixed effects from Experiment 1, for both orders of presentation of phases. Each coefficient represents the estimated number of extra states a phase requires in that condition. Phases 1:1 and 2:2 are matching phaes. Phase 1:2 is mismatching.

The coefficients associated with predictors reveal a somewhat complex picture (Figure 5). Considering that the coefficients indicate the increase or decrease in the number of states required in each condition to achieve the same recognition score compared to the baseline, the coefficients suggest:

- There is a clear distinction between different orderings 1:1, 1:2, 2:2 (interrupted) and 1:1, 2:2, 1:2 (consecutive). The required number of states is minimised for the consecutive ordering
- For either ordering, the need for any more or fewer states when moving from the second phase to the third phase is insignificant.
- Whether the second phase requires more or fewer states than the first depends on whether the second is a match or a mismatch.

Our results cannot confirm the prediction that mismatching phases would require fewer HMM states. It seems that our prediction only holds for the interrupted ordering where there is a monotonic (but not necessarily significant) increase in the number of states required.

If the matching phases are consecutive (1:1, 2:2, 1:2), this seems to help all future phases to reduce the number of required states compared to the first phase (although only the difference between the first and the third phases is significant). However, if the matching phases are interrupted by the mismatching phase (1:1, 1:2, 2:2), every phase requires more states than the one it follows (both second and third phases require significantly more states than the first). This different behaviour based on ordering is visible in the how the coefficients for phases 1:2 and 2:2 have markedly different values in the left and right panels of figure 5. Strikingly, the phase that required the least number of states across all data seems to be phase 2:2 presented as the second phase. This is despite phase 2:2 mapping on to a meaning space twice as large as 1:1.

Order of presentation causing participants to break strategy has an effect beyond whether or not a phase is mismatching. For instance, in the ordering 1:1, 1:2, 2:2, the participant could simply ignore the additional dimension on the final phase to perform at least as well as the second phase, yet there is an (insignificant) increase in the coefficient in the 2:2 phase. Interestingly, the opposite trend can be seen in the other ordering, where changing over to a mismatching phase results in an (insignificant) increase in the number of states required.

## Experiment 2

Experiment 1 provided important evidence of the effects of matching and mismatching signal and meaning space topologies. When there is a one to one mapping between signal and meaning spaces, participants tend to take advantage of it. Indeed, even in our conditions designed to produce a dimensionality mismatch, participants used duration as another signal dimension. Despite this, we were still able to find significant effects of the matching phases compared to the mismatching phases on the amount of movement in signals, the consistency of iconic strategies and how predictable recognition mistakes were.

Experiment 2 was a very similar signal creation experiment. It tested the same hypothesis as Experiment 1, but the design was altered to counter two possible problems with Experiment 1:

1) Duration was used as a dimension by some participants, meaning there wasn't really a "mismatch" even with the 1:2 phase.

2) Participants created signals for a very small meaning set in Experiment 1 (5 or 9 meanings depending on the phase), which was seen in its entirety before the experiment. This made it easier for participants to create a completely holistic signal set without the need for structure. Only one participant treated meanings holistically in Experiment 1 (using frequencies of pitch contours to differentiate meanings). However, we feel that this is still a flaw in the experimental design, as this strategy would soon become maladaptive as meaning numbers rise. In the real world, continuous meaning dimensions are much more nuanced than only having 3 or 5 gradations.

To counter these problems, two alterations have been made in Experiment 2:

1) Phase 1:2 in Experiment 2 has been dubbed a "match" phase, and a new phase 1:3 has been instated to be sure there is a dimensionality mismatch.

2) Participants do not create signals for every possible meaning, but a subset of them. This is explained further in the *Meanings* section below.

## Methods

**Participants.** Participants were recruited at the VUB in Brussels. 25 participants took part in the experiment; 8 male and 17 female. Participants had an average age of 21 (*SD* = 3.2). As in Experiment 1, we asked participants to list the languages they speak, with level of fluency, and to self-report their musical proficiency (on a scale of 1-5).

**Signals.** As in the first experiment, there was a continuous signal space built using the *Leap Motion* sensor to convert hand motion into sounds. However, in this experiment, signals could only be manipulated in pitch. Participants manipulated the pitch in the same way as in Experiment 1, along the horizontal axis. There was an exponential relationship between hand position co-ordinates and signal frequency. The vertical axis was not used at all in this experiment, meaning that, including duration, the number of signal dimensions could not be more than 2. However, participants were not explicitly told to use duration in order to make the results from Experiment 1 more comparable with Experiment 2. Again, participants were given clear instructions on how to use the sensor, and were given a practice period to get used to the mapping between the position of their hand and the audio feedback before the experiment started.

**Meanings.** The meaning space again consisted of a set of squares, but in this experiment they differed along three continuous dimensions: size, shade of orange, and shade of grey. Squares differed along different numbers of dimensions in each phase (Figure 6). In contrast to the first experiment, participants only saw a subset of the possible meanings. Each dimension was divided into 6 gradations, meaning that the meaning space grew exponentially with the number of dimensions (see description of phases below). Having 6 gradations of difference on meaning-space dimensions meant the meaning space is big enough to have make productive systems useful, but coarsely grained enough to not make the discrimination task impossible. Further to the reasons given above, this aspect of the experimental design made an incentive for participants to create productive systems that extend to meanings they have not seen. The subset the meanings participants saw were randomly selected, but participants were explicitly told about all of the possible dimensions. This pressure to make productive systems because one has only seen a subset of a bigger meaning space has been demonstrated in experiments such as Kirby et al. (2008) and Kirby, Tamariz, Cornish, and Smith (2015).

Two of the meaning dimensions in this experiment were “shade of grey" and “shade of orange". In pilot studies, we originally had the squares differ in shade of orange (which we controlled using the RGB ratio of green to red) and the brightness value. However, this made the squares at the darker and redder end of the scale very difficult for participants to tell apart, as they all appeared the same dark brown colour. To solve this, we used striped squares with alternating grey and orange stripes (see figure 6). This gives the same effect of squares differing in shade of orange and brightness, but squares at both ends of the spectrum can be distinguished just as easily.

**Procedure.** The procedure in Experiment 2 was nearly the same as Experiment 1. There were still 3 phases, each with a practice round and an experimental round, which were both the same. Each round has a signal creation task and a signal recognition task. However, the phases were slightly different.

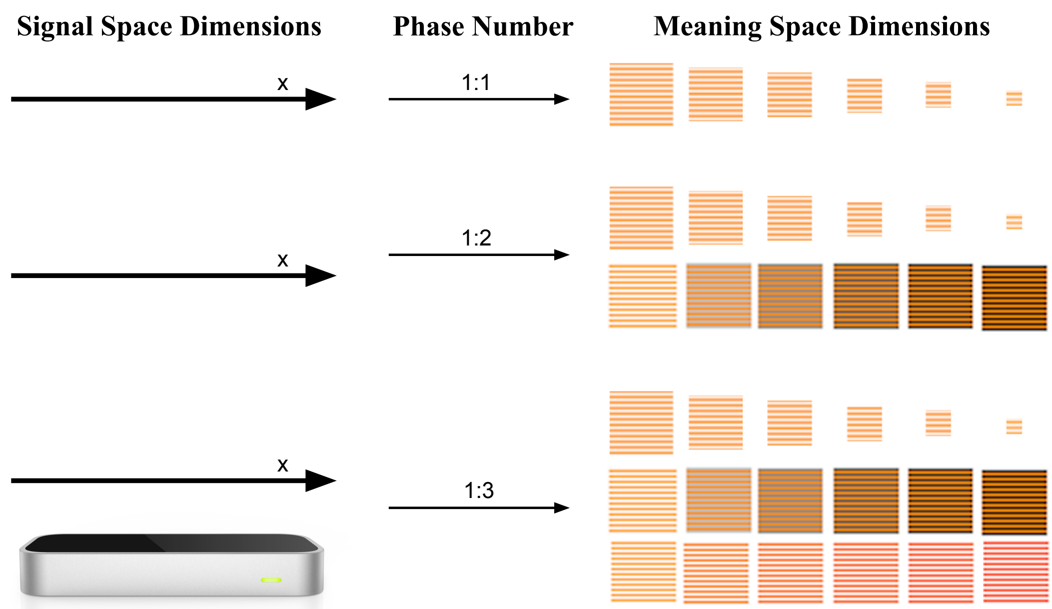


*Figure 6*. The signal and meaning dimensions used in experiment 2 in each of the 3 phases.

**Phases.** All participants had phases presented in the same order: 1:1, 1:2, 1:3. The "1" here refers to 1 signal dimension (pitch), in order to make these phase labels consistent with the phases in Experiment 1. However, since we have learnt to expect participants to use duration as a signal dimension, it is important to remember that the meaning dimensions only outnumber the signal dimensions in a meaningful way in phase 1:3.

**Phase 1:1.** In phase 1:1, there were 6 squares that differed in 6 gradations of size. All 6 squares were presented in a random order.

**Phase 1:2.** In phase 1:2, there were 36 possible meanings. Meanings differed along two dimensions, 6 gradations of size and 6 shades of grey stripes (See Figure 6.) 12 meanings were chosen at random from this set of 36. Participants were then presented with them in a random order. Participants were explicitly told about the introduction of the new meaning dimension at the beginning of the phase.

**Phase 1:3.** In phase 1:3, participants were presented with 12 squares in a random order that differed along three dimensions, 6 gradations of size, 6 shades of grey stripes and 6 shades of orange stripes (See Figure 6.) This made a possible number of 216 squares, which were chosen from at random. This does mean that some participants saw more “evidence" of some dimensions than others in the subset of squares that they saw. However, as with phase 1:2, all participants were explicitly told about the introduction of the third meaning dimension at the beginning of the phase.

**Signal Recognition task.** As in the first experiment, participants completed a signal recognition task. They heard a signal they had created, and were asked to identify its referent from an array of three randomly selected squares from the set of possible squares in the current phase, plus the correct referent, so four squares in total. They were given immediate feedback about whether they were correct, and if not, what the correct square had been. Their performance in this task was recorded for use in the analysis. The distance in the meaning space they were from the correct answer was also recorded in the same way that it was in Experiment 1.

**Post-experimental questionnaire.** The questionnaire asked about the strategies that the participant adopted during each phase of the experiment. As in the first experiment, the questionnaire was free-form. Participants were also asked to name the 6 shades of orange used in the experiment, in order to see if they did indeed label them all "orange", and to see if and how they categorised the colours affected their signals. The shades used in the experiment had been designed to all be perceived as orange. Only 17 participants completed this later part of the questionnaire because of experimenter error.

## Results

### Signal Creation Task

### Descriptive Statistics

In this experiment, signals were on average 2.3 seconds (approx. 252 frames long). The average duration of signals rose by about 20 frames each phase ($\chi^2(1) = 7.9$, $p < 0.005$).

As in Experiment 1, meaning dimensions were coded to reflect the continuous way they differed, i.e. the smallest square was coded as having the value of 1 for size, and the biggest square a value of 6, while the lightest grey/orange stripes were given a value of 1 for shade/colour, and the darkest had a value of 6. Again, across all phases, the size of square was the best predictor for the duration of the signal ($\chi^2(1) = 63.3$,

$p < 0.001$), with signals for the smallest squares having a mean duration of 1.55 seconds (*SD* = 1.26s), and the largest squares having a mean duration of 2.7 seconds (*SD* = 1.9s). However, in this experiment size was also the best predictor for the mean pitch of the signals ($\chi^2(1) = 15.7$, $p < 0.001$). The smallest squares had a mean pitch of 403Hz, and the largest squares had a mean pitch of 333Hz. Again, we take this as evidence for the use of relative iconicity.

We again looked at the standard deviations of individual signal trajectories to see if the degree of mismatch in the signals affected the amount of movement in the signals. There was no significant difference between the two matching phases (Phases 1:1 and 1:2), in fact, the mean standard deviation in these phases was nearly identical (around 28mm, *SD* = 31.5). However, the SDs from phase 1:3, the mismatch phase, was significantly higher (*M* = 33.8mm, *SD* = 34.4) than in the other two phases ($\chi^2(1) = 6.9$, $p < 0.01$) indicating more movement in the mismatch phase. Figure 7 shows how this effect manifested itself in the signals of one participant where the differences between phases were particularly marked.

## Predictability of signal trajectories

We again calculated the predictability values for each of the signal trajectories in a repertoire in the same way as we did in Experiment 1. We were interested to see if whether a phase was matching or mismatching had an effect on how predictable the signals were. Using a linear mixed effects model and controlling for duration and participant number as a random effect, and size of square as a fixed effect, we found that whether the signal was produced in a matching phase or not correlated with how predictable a trajectory was ($\chi^2(1) = 11.2$, $p < 0.001$). The value was closer to 0 (so more predictable) in phase 1:1 (*M* = 95), and got less predictable with each phase (phase 1:2 *M* = 119, phase 1:3 *M* = 145).

## Signal Recognition Task

We used a linear model to test if musical proficiency predicted performance in the signal recognition task, and, as in Experiment 1, found that it did not ($F(1,23) = 0.03, p = 0.28$).

Overall, participants were slightly worse at recognising their own signals in Experiment 2 than in

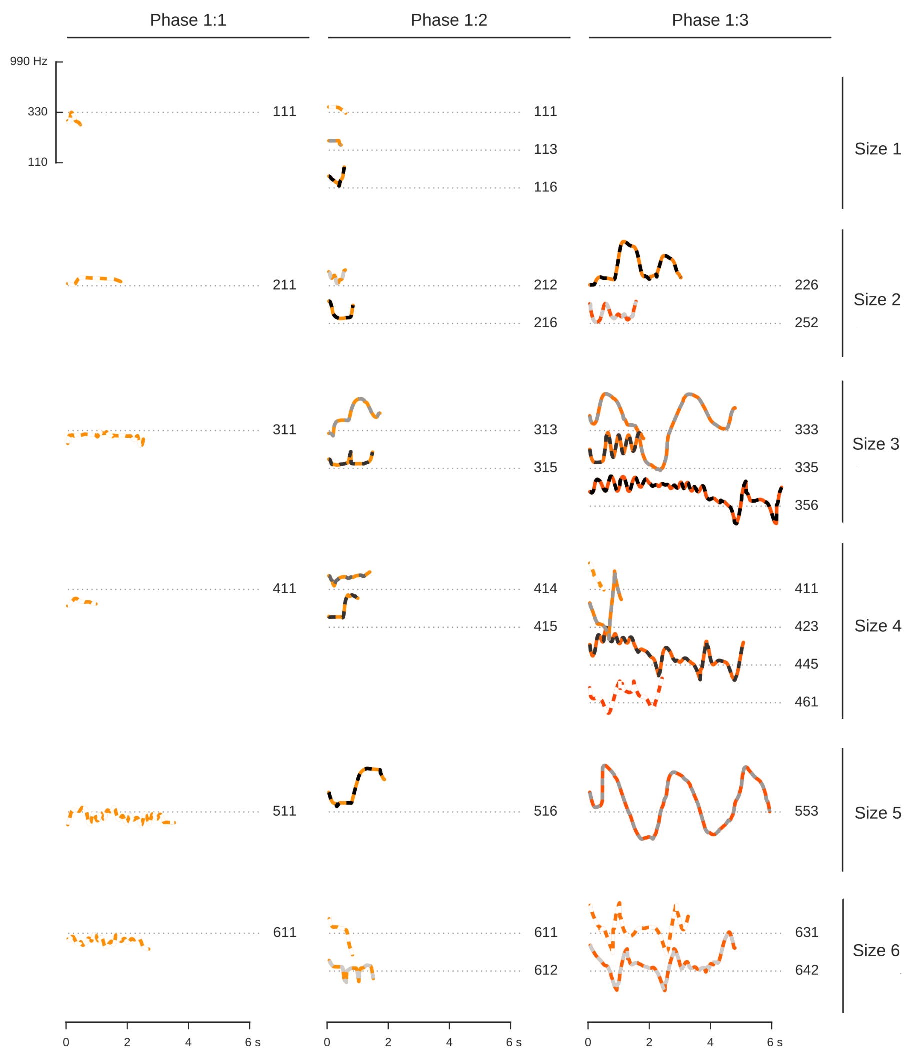


*Figure 7*. The entire signal repertoire of one participant in all three phases. The colour of the stripes in the pitch tracks represents the colours of the squares they represent. Square size is denoted along the right-hand side. The numbers by each pitch track are the file names of each meaning which also encode the size and shade of orange and grey. Signals produced in phase 1:3 have visibly more movement than in the other two phases.

Experiment 1. They recognised their signals with a mean of 56% correct (*SD* = 13%), again with a chance level of 25%. Using a linear model, we tested whether participants improved in their performance throughout the experiment, as they did in Experiment 1, but found no correlation ($F(1,23) = 1.39$, $p = 0.24$). Success stayed constant across phases around the 56% mean. The lack of improvement as participants became more experienced was probably because the meaning space expanded so rapidly with each phase, making the recognition task much more difficult.

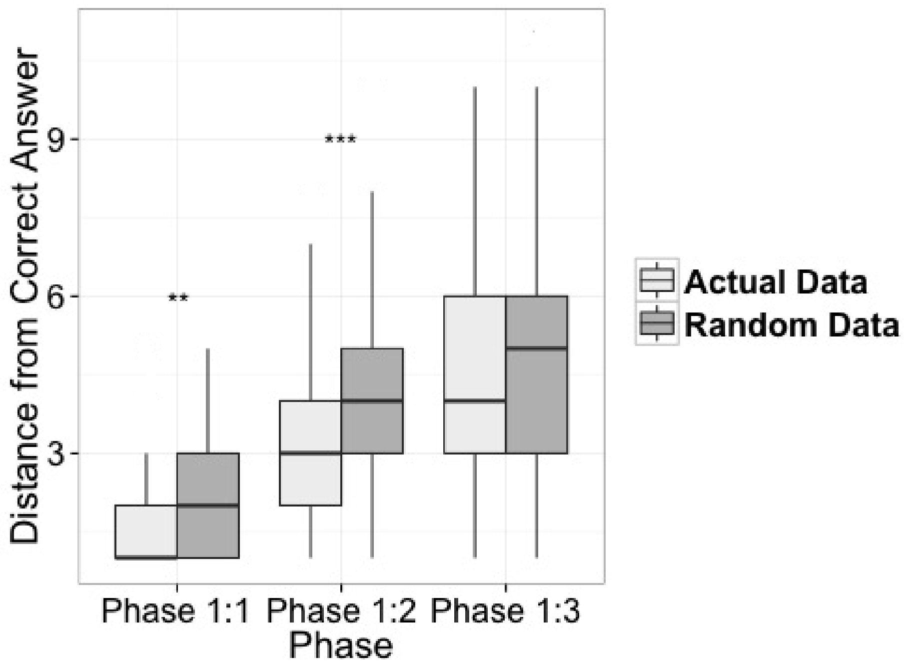


*Figure 8*. A graph showing the distance from the correct answer participants were in each phase when choosing incorrectly in the signal recognition task.

Again, when participants were incorrect, we were able to measure the distance between their answer and the correct answer. We did this in the same way as we did in experiment 1. Using a mixed effect linear model, and controlling for participant number as a random effect and square number as a fixed effect, we found that with incorrect choices produced across phases, participants were closer to the correct square (*M* = 3.3 steps away, *SD* = 2) than if they had chosen at random (*M* = 4 steps away, *SD* = 2.1) ($\chi^2(1) = 22.4$, $p < 0.001$) (see figure 8), the difference between actual and random data was significant within phases 1:2

and 1:3 as well.

In later phases, incorrect distances were higher because of the bigger meaning space. Therefore, 4 meanings chosen at random would have a much bigger mean distance between them in the bigger meaning spaces. As a result, comparison between phases of distance from the correct answer is not indicative of participants having problems. However, bigger effect sizes when comparing the actual data with random data might indicate more reliance on iconicity. This is because choosing meanings close to the correct meaning indicates use of iconicity. When there is no iconicity, the answers should be more similar to the random data. The effect size for the comparison between the actual data and the random data in phase 1:3 was smaller ($d_r = 0.27$) than in the other two phases ($d_r = 0.46$), suggesting that in phase 1:3 there was less reliance on iconic strategies.

**Post-experimental questionnaire**

In Experiment 2, every participant had a strategy. Generally, participants in Experiment 2 reported the experiment to be more difficult than participants in the first. In phase 1:1, participants encoded size directly with pitch or duration (80% self-reported). Participants tended to stick with the same strategy for size, but developed strategies on top of that to cope with the different shade elements, and by phase 1:3, 56% of participants self-reported using a strategy that relied on movement, patterns or pattern frequencies.

Responses to the colour categorisation part of the questionnaire were very variable, ranging from 2-6 categories over the 6 squares, with a mean value of 4.2 categories, though most categories included the word orange, such as “light orange", “dark orange", “red orange", “sunset orange", “blood orange", but people also labelled the darkest shade “red". There was no interaction between the number of categories that participants separated the squares into and how well they did in phase 1:3, which was the only phase to use different shades of orange ($F(1,16) = 1.56$, $p = 0.23$).

## Hidden Markov Models

The data from the second experiment were processed identically to the first from continuous trajectories to HMMs. Then, the number of states for the HMMs, i.e. the best predictor from the first experiment, was used to predict the recognition scores using a linear mixed effects model while controlling for participant number and phase.

The second experiment did not yield the same results as the first one. The regression did not predict recognition scores using the number of states in any representation of the signals. Further analysis was performed to see if any of the other candidate predictors worked for this particular data set, but no predictor performed well. In other words, we failed to demonstrate that the HMM models captured participant performance for this experiment.

To investigate which aspect of the second experiment was different, we modelled a third data set from Little et al. (2015), summarised in the introduction of the current paper. The only difference between the experiment presented in Little et al. (2015) and Experiment 1 is that the former used *discrete* meanings that don't have an intuitive, natural ordering, such as various textures or colours. This prevented the participants from exploiting the natural ordering of a continuous meaning space as they do in the current experiments, but retains any dimensionality effects.

We modelled this data set using HMMs and analysed it in the same way as Experiment 1. The fixed effect coefficients show that ordering of phases is still important for the discrete case (see Figure 9). While for both orderings, the 2:2 phase requires more states than the 1:2 phase, this difference is only significant in the cases where there is no strategy change necessary (with interrupted order). This shows that the continuous data set is more efficiently represented using relative iconicity that doesn't change across the experiment, whereas the discrete data set is most efficiently represented in the mismatching phase, but only after a strategy within a matching phase has been established first. This demonstrates that the types of meanings do modulate the efficiency of iconic and non-iconic strategies, where more continuous, ordered meaning spaces are better represented using relative iconicity.

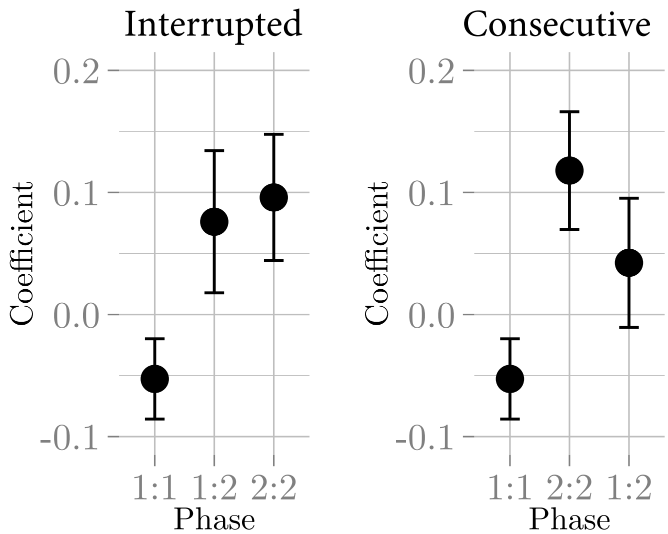


*Figure 9.* Fixed effects from the discrete case for both orders of presentation of phases, covered in littlelinguistic. Each coefficient represents the estimated number of extra states a phase requires in that condition.

The analysis of the data from Little et al. (2015) adds to our information, giving us knowledge of how the model behaves using data from three different experiments. The HMMs make reasonable predictions about participant behaviour in Experiment 1 and in Little et al. (2015). This raises the question of what causes the issue with Experiment 2. The most salient different between the two experiments was the absence of a two-dimensional signal space in Experiment 2, as only pitch was used, as well as the 1:3 phase. Accounting for what exactly would cause HHMs to not be able to model this data in an intuitive way is not clear. Despite this, we think that HMMs are a very worthwhile method to pursue, illustrated by where we have succeeded. However, further work needs to focus on refining our understanding of what predictions make sense for different data sets.

## Discussion

We set out to experimentally investigate two hypotheses:

1) When the topologies of signal and meaning spaces are the same, this facilitates the emergence of iconic signals.

2) When the number of meaning dimensions outnumbers the signal dimensions, this facilitates the

emergence of combinatorial structure.

In both experiments, we found correlation between the structure of signal repertoires and the structure of the meaning space, indicating a prevalence of relative iconicity. This was particularly marked when signal and meaning spaces had the same number of dimensions. We also found evidence for more movement in signals in phases where there was a mismatch between signal and meaning spaces, suggesting a departure from relative iconicity to a possibly more structured signalling system. Signals were also longer in later phases in Experiment 2, which perhaps points to more sequential encoding. Lewis and Frank (2016) previously showed that longer word forms are associated with meanings with more complexity, and signal duration has also been used as a measure for complexity in experimental studies such as Roberts et al. (2015).

During phases with matching dimensionalities, participants produced signals that were more predictable, given a participant's entire repertoire, than signals produced within mismatching phases. This is probably due to the mismatching phases producing signals with more movement, which is less predictable than static signals indicative of relative iconicity. We also found that in matching phases, when participants were incorrect, they were more likely to choose meanings that were closer to the correct meaning than if they had chosen at random, again suggesting a reliance on relative iconic strategies.

The above results provide evidence for the first hypothesis, that matching topologies incentivise participants to produce signals with relative iconicity. They also show that more movement and complexity was present when meaning dimensions outnumbered signal dimensions. However, exactly how we can characterise this movement remains unclear. One possibility is that the movement in our signals is iconic, for instance, representing the stripes of meanings in Experiment 2. However, the post-experimental questionnaires do not support this narrative. It is clear from the questionnaires that participants often used structural strategies, in that specific elements or dimensions of the signal refer to different dimensions of the meaning that are then combined to refer to the whole meaning. However, structure such as this is not indicative of combinatorial structure as we defined it in the introduction. That is, the building blocks are not meaningless but correspond to dimensions in the meaning space. However, there is very little flexibility in the way signal dimensions can be combined in our experiments, and parts of the signals/meanings cannot occur in isolation (that is, every signal has to have both a pitch and a duration). In this respect, the structure is neither combinatorial nor compositional but something in between, and possibly something that could be reanalysed by speakers to be combinatorial structure through the mechanisms proposed by Goldin-Meadow and McNeill (1999). Investigating what might cause this reanalysis to happen would make a good departure for future experimental work, perhaps having participants creating signals for bigger and less structured meaning spaces to get rid of the inhibiting effects of iconicity.

Further to the above, we also gathered evidence about structure in our signals using Hidden Markov Models. We found interaction between number of states, phase, and phase order in Experiment 1, but were not successful in doing this for Experiment 2. Despite this, we feel that with some fine-tuning Hidden Markov Models will be a worthwhile tool for measuring combinatorial structure in artificial signalling experiments in the future.

## Further Work

One of the major difficulties we faced in the analysis of this experiment was variation in participants' behaviour. In a population of signallers, especially without iconicity, diversity of signalling strategies is not beneficial, as signallers need to settle on a shared strategy to be mutually understandable. In order to address this problem, our next step will be to develop this paradigm with social coordination experiments where pairs or groups of participants create shared communication systems. A communication game will also allow us to identify effects that are the result of interaction as opposed to the pressure for expressivity on its own.

Another next step will lie in the extension of the paradigm to look at other ways to manipulate the mappability between signal and meaning spaces. In the current experiments, participants were describing a continuous ordered meaning space with a continuous signal space. Further, as the meaning space in our experiment was very structured, what we found was signal structure that directly corresponded to the structure in the meaning space. However, having meaning space dimensions that are not continuous will obfuscate the signal-meaning mapping in a way that will make iconic strategies much more difficult. Work in this area has already started (Little et al., 2015), but we are still pursuing research on how different meaning spaces can affect the emergence of signal structure on different levels. In this vein, we have also run further experiments with less internal structure in the meaning space in order to obtain signals that have structure more analogous to phonological structure than compositional structure (Little, Eryılmaz, & Boer, in press).

Finally, progressively more advanced Hidden Markov Model variants can be employed where the Markovian assumption is relaxed. This will both enable using new dimension types, such as duration, in the HMMs, and also potentially provide more theoretically justified model selection criteria, such as the implicit selection of the number of states in Dirichlet Process HMMs.

## Conclusion

In conclusion, we have shown that the topology and dimensionality of a signal space will affect the emergence of structure and iconicity: the more closely the topologies of the signal and meaning space correspond, the easier it is to use iconic structure. If there is no good correspondence, we see more movement in the signals: perhaps the first steps towards structure (either combinatorial or compositional). These findings are important to understand how linguistic modality affects the emergence of structure in real world languages. The manual modality has more signal space dimensions than speech. This may help explain why some emerging sign languages go through a phase where they do not appear to use combinatorial structure, but do use iconicity extensively. Our experimental results indicate that having more dimensions will not only affect how quickly the signal-space gets overcrowded, but also to what extent signalling strategies that use relative iconicity can be used. It is for these reasons, we would like to argue that our two hypotheses are intrinsically linked as they are both tied up in the topology and dimensionality of the signal space.

As a final point, our results are also important for researchers conducting artificial language experiments with signal-space proxies. The topology of the signal space being used has significant effects on the iconicity and structure which emerges in the experiment which researchers need to be mindful of. Importantly, understanding these effects, as we have attempted to do here, will put us in a better position to separate the effects of signal space topology from other effects under investigation in the broader literature.

**Acknowledgements**

Thanks to 3 anonymous reviewers and the editor for their valuable comments and suggestions. Thank you to Catriona Silvey for very thorough comments on a draft of this paper. Thank you to Yannick Jadoul for the improvements made to figure 7. The project was financed by the ERC starting grant project ABACUS, grant number 283435.

References

Baker, J. M., Deng, L., Glass, J., Khudanpur, S., Lee, C.-H., Morgan, N., et al. (2009). Developments and directions in speech recognition and understanding, part 1 [dsp education]. *Signal Processing Magazine, IEEE*, *26*(3), 75–80.

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, 201605782.

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

Crasborn, O., Hulst, H. van der, & Kooij, E. van de. (2002). Phonetic and phonological distinctions in sign languages. In *A paper presented at intersign: Workshop* (Vol. 2).

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

Eryılmaz, K., & Little, H. (2016). Using leap motion to investigate the emergence of structure in speech and language. *Behavioral Research Methods*.

Fay, N., Lister, C. J., Ellison, T. M., & Goldin-Meadow, S. (2014). Creating a communication system from scratch: gesture beats vocalization hands down. *Frontiers in Psychology*, *5*, 354.

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

Goldin-Meadow, S., & McNeill, D. (1999). *The role of gesture and mimetic representation in making language the province of speech*. na.

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

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.

Lewis, M. L., & Frank, M. C. (2016). The length of words reflects their conceptual complexity. *Cognition*, *153*, 182–195.

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)* (Vol. 10, pp. 479–481). World Scientific.

Little, H., Eryılmaz, K., & Boer, B. de. (in press). Conventionalisation and discrimination as competing pressures on continuous speech-like signals. *Interaction Studies*.

Little, H., Eryılmaz, K., & de Boer, B. (2015). Linguistic modality affects the creation of structure and iconicity in signals. In D. C. Noelle et al. (Eds.), *The 37th annual meeting of the cognitive science society (cogsci 2015)* (pp. 1392–1398). Austin, TX: Cognitive Science Society.

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language. *Philosophical Transactions of the Royal Society B*.

Nowak, M. A., Krakauer, D. C., & Dress, A. (1999). An error limit for the evolution of language. *Proceedings of the Royal Society of London B: Biological Sciences*, *266*(1433), 2131–2136.

Räsänen, O. J., Laine, U. K., & Altosaar, T. (2009). A noise robust method for pattern discovery in quantized time series: the concept matrix approach. In *Interspeech* (pp. 3035–3038).

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.

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.

Schliep, A., Georgi, B., Rungsarityotin, W., Costa, I., & Schonhuth, A. (2004). The general hidden markov model library: Analyzing systems with unobservable states. *Proceedings of the Heinz-billing-price*, *2004*, 121–135.

Schwarz, G., et al. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

Scott-Phillips, T. C., & Blythe, R. A. (2013). Why is combinatorial communication rare in the natural world, and why is language an exception to this trend? *Journal of The Royal Society Interface*, *10*(88), 20130520.

Verhoef, T., Kirby, S., & Boer, B. (2015). Iconicity and the emergence of combinatorial structure in language. *Cognitive science*.

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*, 57–68.

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

Zuidema, W., & Westermann, G. (2003). Evolution of an optimal lexicon under constraints from embodiment. *Artificial Life*, *9*(4), 387–402.

# Appendix A

## HMMs

The HMMs used in this study are HMMs with multivariate continuous Gaussian emissions, using the standard Baum-Welch algorithm for unsupervised training. We used a slightly modified version of the Python wrappers for GHMM library as our HMM implementation (Schliep, Georgi, Rungsarityotin, Costa, & Schonhuth, 2004).

Since Baum-Welch is an expectation-maximisation algorithm, it is susceptible to getting stuck in locally optimal solutions. To overcome this, for each combination of parameters, we randomly initialise multiple models, and pick the one with the highest likelihood. We chose to compare 100 random initialisations for each parameter set.

### Model Selection

The parameter for the number of hidden states is the only one not estimated by the Baum-Welch algorithm. It also determines the size of the model since each additional state adds new parameters to the model. We have to perform model selection over candidate models to approximate the best number of states for each dataset. We do this by comparing the Bayesian Information Score (BIC) values of the competing models, picking the one with the lowest BIC. BIC is a measure that balances the likelihood of the model and the size of the model, providing a model with both a high likelihood and a minimal size (Schwarz et al., 1978).

### Training data

For each HMM, the training data consists of all the signal data from a particular participant at a particular round. Since there are three possible data projections, three models are trained per parameter set. In each phase, there are 5 to 12 signals (depending on the specific phase and experiment), and all of them are used for training (since this is already quite a small amount of data to train these models on). The same BIC selection procedure is used to pick the best projection.

The number of states varied between 2 and 30. The number 30 is an upper limit inspired by the number of states that would be needed if there were one state per meaning in two dimensions ($12 \times 2 = 24$), an inefficient, one-to-one, iconic encoding. The BIC usually stops decreasing significantly after this point as well, and training larger models becomes increasingly time consuming, so we capped this parameter at 30.

In total, these add up to $(30-2) \times 100 = 2800$ HMMs trained per projection per phase per participant, of which the one with the lowest BIC score is used as the best model. Each phase for each participant was modelled by exactly three HMMs, one for each projection. The best projection for each experiment was chosen using the mixed effects regression outlined in *Appendix B*.

**Algorithm 1** HMM training and selection for each projection

```
function FITHMM(trajectories)
    hmm ← nil
    bic ← 999999
    nStates ← 2
    maxStates ← 30
    while nStates ≤ maxStates do
        for 1 : 100 do
            hmm′ ←HMM(nStates)
            for trajectory in trajectories do
                hmm′ ←BAUMWELCH(hmm’, trajectory)
            if BIC(hmm′) < bic then
                hmm ← hmm′
                bic ←BIC(hmm’)
        nStates ← nStates + 1
    return hmm
function ANALYZEDATA(participants, data)
    models ← {}
    for pr in participants do
        for phase in 1:3 do
            trajectories ← data[pr][phase]
            models[pr][phase] ←FITHMM(trajectories)
    return models
```

## Appendix B

### Mixed Effect Linear Regression

Let $O(p) \in \{1,2,3\}$ be the order of phase $p$. Then the regression equation can be expressed as:

$$\begin{aligned} \text{score}_{id,p} &= \alpha_0 + \alpha_{id} + \varepsilon \\ &+ N_{states} \times slope(p) \end{aligned} \tag{B-1}$$

where

$$slope(p) = \begin{cases} \beta_1 & \text{if } p = 1:1 \\ \beta_2 & \text{if } p = 1:2 \;\&\; O(p) = 2 \\ \beta_3 & \text{if } p = 1:2 \;\&\; O(p) = 3 \\ \beta_4 & \text{if } p = 2:2 \;\&\; O(p) = 2 \\ \beta_5 & \text{if } p = 2:2 \;\&\; O(p) = 3 \end{cases} \tag{B-2}$$

The coefficient values were calculated as $\alpha_0 = 0.640$, $\beta_1 = 0.077$, $\beta_2 = 0.193$, $\beta_3 = 0.053$, $\beta_4 = 0.000$, and $\beta_5 = 0.241$, where $\alpha_0, \beta_1, \beta_2, \beta 5$ are found to be the predictors for which $p < 0.05$. The $\alpha_{id}$ intercepts for each participant varied in the range $[-0.237, 0.189]$.

On Figures 5 and 9, the coefficients plotted as *Ordering 1* are $\beta_1$, $\beta_2$, $\beta_5$, and the ones plotted as *Ordering 2* are $\beta_1$, $\beta_4$, $\beta_3$.