The cumulative cultural evolution of category structure in an infinite meaning space

Jon W. Carr, M.A. (Hons)

August 2013

Submitted in partial fulfilment of the degree of Master of Science to

School of Philosophy, Psychology and Language Sciences
The University of Edinburgh

Abstract

Current iterated learning experiments use meaning spaces that are discrete, finite, pre-specified, and low-dimensional. Such meaning spaces are poor representations of the world. For this reason, we have conducted two experiments to look at the cumulative cultural evolution of category structure in an infinite meaning space.

In the first experiment, the number of words used to describe the stimuli collapses dramatically after only a few generations. Within a few more generations, a system emerges that arbitrarily divides the space into a small number of categories pertaining primarily to the size and shape of the stimuli. In the second experiment, we apply an artificial constraint which prevents the size of the languages from collapsing. This constraint was implemented to model the pressure for expressivity that exists in languages when they are used functionally for communication. We predicted that this would allow compositional structure to emerge so that the space could be carved up in more finely grained and/or higher dimensional ways using a compressible linguistic system. However, there was little sign of compositionality emerging under the parameters of this experiment.

Although the meaning space presented here is a simple one, we hope that this project represents a first step towards thinking about how iterated learning experiments deal with the problem of discrete infinity. We briefly discuss the background literature, then present the methods and results for this project, and end with some discussion about how the results relate to our research questions.

Acknowledgements

First I must thank my supervisors, Simon Kirby and Hannah Cornish. When Hannah mentioned the words 'infinite meaning space', I knew immediately that I wanted to be involved. She has always been super enthusiastic about the project and ready with sound advice. I owe you many Cherry Colas! Simon is never short of ideas and has a wonderful clarity of mind. He too has contributed to this greatly by being able to envision the end result before I had even grasped the nuts and bolts. Nevertheless, all errors in what follows are my own.

Kenny Smith has been a top-notch programme director during my two years on the MSc. Reliable, funny, and just generally as cool as a cucumber. And the great thing is, he really does care. Cheers!

The Carnegie Trust contributed half of my tuition fees, so I definitely owe them a huge thank you. Without their investment, this handful of pages would never have existed. I also want to thank my fellow students on the programme, as well as the other members of the LEC research group, who have made these two years as enjoyable as they have enlightening.

Too many other people to list here have contributed in so many other ways – you know who you are – but most important of all is Wil, who has shown me support like no one else. You are my *sine qua non*.

Now to the task at hand.

Contents

List of figures

List of tables

1. Introduction

This chapter provides a short introduction to the concepts that will be discussed in this dissertation. Section 1.1 covers general explanations for the evolution of language. Section 1.2 describes the iterated learning paradigm. Section 1.3 discusses meaning spaces and categorical perception. Section 1.4 reflects on the meaning spaces used in previous experiments. Section 1.5 summarizes the research questions and hypotheses that will be explored in this project.

1.1. Explanations for language

Language is arguably the defining characteristic of our species, and asking where language comes from is deeply connected to questions about what it means to be human. In their classic paper, Pinker and Bloom framed the evolution of language in terms of a 'conventional neo-Darwinian process' (1990, p. 707). This seems natural given that Darwin's (1859) theory of natural selection has been highly successful in explaining the complexity we observe in nature. However, natural selection has far-reaching consequences beyond biology alone. Human cultures also adapt and evolve, providing us with two modes of inheritance: the genetic and the cultural (Mesoudi, 2011; Richerson & Boyd, 2005). Indeed, Maynard Smith and Szathmáry (1995) cite language and culture as the most recent in a series of evolutionary transitions that have irrevocably changed the way information is stored and transmitted. In addition, the evidence for an innate, domain-specific language faculty (supported by e.g. Chomsky, 1965; Jackendoff, 2002; Pinker, 1994) has been questioned from the perspectives of language acquisition (Pullum & Scholz, 2002), neuroscience (Schoenemann, 2009), and typology (Evans & Levinson, 2009).

Since the publication of Pinker and Bloom's (1990) paper, there has been rising interest in the evolution of language, and various alternative scenarios have been proposed that place greater emphasis on cultural evolutionary dynamics (e.g. Christiansen & Chater, 2008; Croft, 2000; Kirby, 1999). Furthermore, new tools and methods have made the scientific investigation of such dynamics possible (see Caldwell & Millen, 2008; Mesoudi & Whiten, 2008; Scott-Phillips & Kirby, 2010 for reviews). These methods have included

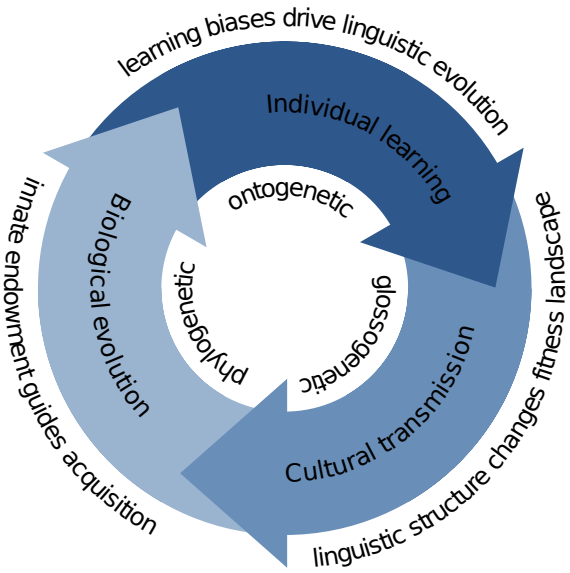


Figure 1.1 Diagram illustrating the interaction between three complex adaptive systems working on three timescales: biological evolution, cultural evolution, and learning. Diagram based on a combination of Figure 1 in Kirby (2002b, p. 189), Figure 1 in Kirby et al. (2007, p. 5242) and Figure 4 in Smith et al. (2003, p. 541).

mathematical (e.g. Griffiths & Kalish, 2007), computational (e.g. Oliphant, 1996), and experimental (e.g. Garrod, Fay, Lee, Oberlander, & MacLeod, 2007) models, as well as the exploration of large datasets (e.g. Lupyan & Dale, 2010). In particular, cultural evolutionary approaches have shown that fundamental properties of language can emerge through social negotiation (e.g. Scott-Phillips, Kirby, & Ritchie, 2009; Steels, 1997; Zuidema & de Boer, 2009) and cultural transmission (e.g. Kirby, Cornish, & Smith, 2008; Kirby, Dowman, & Griffiths, 2007; Smith, Brighton, & Kirby, 2003).

1.2. Iterated learning

Besides biological and cultural evolution, Kirby and Hurford (2002) argue that (a) a third complex adaptive system, learning, should also be accounted for, and (b) the structural properties of language can be explained by the interactions between these three systems. This is illustrated in a feedback loop in Figure 1.1. Kirby and Hurford (2002) place particular emphasis on the repeated induction and expression of language across generations of language user, a process they term 'iterated learning'.

1.2.1. Theoretical outline

Iterated learning refers to 'a process in which an individual acquires a behavior by observing a similar behavior in an individual who acquired it in the same way' (Kirby et al., 2008, p. 10681). Kirby and Hurford (2002, p. 123) describe four basic components of an

iterated learning model: (1) a meaning space, (2) a signal space, (3) one or more language-learning agents, and (4) one or more language-using adult agents. The first cultural generation of agents is given a set of random mappings between the meaning space and signal space, and they construct hypotheses to explain these mappings. This generation of agents subsequently provides the input for a new generation based on the inferences they have made. The second generation then trains a third generation and so forth until either a specified number of generations has elapsed or the languages stabilize in some way. Crucially, the mappings produced by any given agent will include mappings that the agent never observed in its training input, due to the presence of a bottleneck on transmission[1]. When treated as a complex adaptive system, language can adapt to this bottleneck, as well as to the cognitive biases of its learners (see references in the following sections).

1.2.2. Computer simulations

Early work in iterated learning relied on computational simulations of populations of agents. This body of work showed that linguistic properties can arise epiphenomenally during cultural transmission (e.g. Hurford, 1989; Kirby, 2002a; Smith, 2004). For example, Smith et al. (2003) used agent-based modelling to show that compositionality arises as an adaptation to the poverty of the stimulus, since only structured, generalizable languages are stable under the constraint of a transmission bottleneck.

1.2.3. Experiments with human participants

Computer simulations of iterated learning were criticized on the grounds that the cognitive behaviours of the agents were unrealistic (Bickerton, 2007, p. 522). Responding directly to this criticism, and taking cues from earlier experimental work[2], Kirby et al. (2008) devised an experimental analogue of the iterated learning paradigm using adult human learners. In experiment 1, there was no pressure for participants to be expressive and, as a consequence of this, the languages rapidly collapsed to a few distinct words. Kirby et al. (2008) refer to this as 'systematic underspecification', which represents one way in which languages can circumvent the bottleneck. In experiment 2, the authors implemented a filtering system so that the subset of the language that a participant saw in

[1] A limit on how much information can be transmitted from one generation to the next.

[2] Notably, Bartlett's (1932) work on diffusion chains, as well as more recent work with both non-human animals (e.g. Horner, Whiten, Flynn, & de Waal, 2006; Laland & Williams, 1997) and humans (e.g. Galantucci, 2005; Selten & Warglien, 2007).

training was always composed of a set of unique signals, guiding the languages away from underspecification. Although artificial, this modification was intended as an analogue of the pressure for expressivity that exists in functional languages. In this experiment, small sets of meaningful, recombinable units emerged corresponding to the dimensions of the meaning space. The authors refer to this outcome as 'systematic compositional structure'.

A variety of other iterated learning experiments have since been conducted. Smith and Wonnacott (2010) have used the paradigm to show that the predictability of a grammatical marker increases cumulatively as a language is culturally transmitted, suggesting that iterated learning can offer an explanation for language regularity. Perfors and Navarro (2011) have used a combination of Bayesian modelling and experimental work to show that iterated learning gives rise to signal structures that reflect the structure of the meaning space. Verhoef (2012) has shown that iterated learning can explain the emergence of combinatorial structure in whistled songs.

1.3. Meaning spaces and categorical perception

We have briefly alluded to the concept of meaning spaces above. Here we describe them more formally. Gärdenfors (2000), among others, promotes a general approach to semantics in which meanings are represented as points in a convex 'conceptual space' so that the distance between two points corresponds to their similarity[3]. An individual's conception of the world is modelled in terms of an n-dimensional metric space with points in the space representing objects, and regions of the space representing concepts.

1.3.1. Metric spaces, similarity, and the universal law of generalization

A metric space is a set on which a distance function has been defined between elements of that set (Ó Searcóid, 2007, p. 2). For example, the distance function d between any two elements a and b in the set of real numbers $\mathbb{R}$ could be defined as the absolute value of the difference, i.e. $|b-a|$. In this case, we say that the space $\mathbb{R}$ is endowed with the metric $d(a,b) = |b-a|$, yielding the metric space $(\mathbb{R}, d)$. A distance function d on a set X is valid if, for all a, b, and c in X, four conditions are met: $d(a,b) \geq 0$ (non-negativity); $d(a,b) = 0$ iff $a = b$ (the coincidence axiom); $d(a,b) = d(b,a)$ (symmetry); and $d(a,b) \leq d(a,c) + d(c,b)$ (the triangle inequality). Common examples of distance metrics include the

[3] Gärdenfors does not make the stronger claim that this is how the brain processes meaning, only that it is useful as a model.

Euclidean distance, Hamming distance, and Manhattan distance.

By using a metric space as a model of the underlying physical parameter space, we can proceed to measure the similarity between two meanings by transforming the measure of distance into a measure of similarity. This is typically performed using an 'exponentially decaying function of distance', such as $s_{i,j} = e^{-c \cdot d_{i,j}}$ where c is a sensitivity parameter (Gärdenfors, 2000, p. 20). This is referred to by Shepard (1987) as the 'universal law of generalization'. He argues that generalization 'govern[s] the behaviors of all sentient organisms' (1987, p. 1317), and that the mathematical concept of a metric space has been evolutionarily internalized as the mechanism for generalizing from one situation to another. Regardless of the truth of this, metric spaces provide a useful means for modelling how individuals conceptualize the world.

1.3.2. Categorical perception

Categorical perception was originally intended to explain the categorical nature of speech perception (Liberman, Harris, Hoffman, & Griffith, 1957), but has since been co-opted to explain the discrete perception of continuous sensory phenomena (see Harnad, 1987). It is suggested that categorical perception operates by a process of induction. For example, if we know that x_1 is an element of X, and we know that x_2 is similar to x_1, then it follows that x_2 must also be an element of X. In terms of the psychological spaces we mentioned above, Harnad (1987) describes categorical perception as forming within-category compression (pinching the space) and between-category separation (stretching the space), such that members of a single category tend to be perceived as more similar and members of separate categories tend to be perceived as more different. There is ongoing debate regarding the degree to which categorical perception is innate versus learned and the degree to which it is influenced by language (see Goldstone & Hendrickson, 2009 for some discussion). Aside from speech perception, categorical perception has been widely studied in the perception of colour (e.g. Winawer et al., 2007), faces (e.g. Beale & Keil, 1995), and shape (e.g. Roberson, Davidoff, Shapiro, 2002).

1.4. Meaning spaces in previous iterated learning experiments

Previous iterated learning experiments have relied primarily on meaning spaces that are discrete, finite, pre-specified, and low-dimensional. Kirby has described these as 'fixed, monolithic meaning space[s]' (2007, p. 256). Table 1.1 provides a catalogue of the meaning

spaces used in a selection of previous experiments. This is by no means a complete list, but it highlights the relatively small scale of the meaning spaces that have been used previously.

Table 1.1 Catalogue of meaning spaces used in previous iterated learning experiments

Study	Type of meaning space	Magnitude
Brown (2008)	Discrete	3 dimensions, 27 meanings
Carr (2012)	Discrete	3 dimensions, 27 meanings
Cornish (2011) (experiments II—IV)	Discrete	3 dimensions, 27 meanings
del Giudice (2012) (experiment 2)	Discrete	4 categories, 20 meanings
Fay et al. (2010)	Discrete	4 categories, 20 meanings
King (2011)	Discrete	3 dimensions, 27 meanings
Kirby et al. (2008)	Discrete	3 dimensions, 27 meanings
Matthews (2009)	Continuous	2 dimensions, 100 meanings
Murray, K. (2009)	Discrete	3 dimensions, 27 meanings
Murray, L. (2010)	Discrete	3 dimensions, 27 meanings
Perfors and Navarro (2012)	Continuous	2 dimensions, 36 meanings
Seyfarth (2010)	Discrete	3 dimensions, 27 meanings
Smith and Wonnacott (2010)	Discrete	2 dimensions, 8 meanings
Winters (2009)	Discrete	3 dimensions, 18 meanings

1.4.1. Discrete meaning spaces

The meaning space in Kirby et al. (2008) is three-dimensional with each dimension (colour, shape, and movement) varying over three discrete qualities for 27 meanings. This basic structure has been replicated in several other experiments (e.g. Brown, 2008; Murray, 2009; Winters, 2009) with some variation. To take another example, the space in Smith and Wonnacott (2010) has two discrete dimensions (animal and plurality) for a total of eight meanings (four animals in the singular or plural).

1.4.2. Continuous meaning spaces

Matthews (2009) implemented a continuous meaning space comprising 100 stimuli that were produced by morphing triangles into rectangles in two orientations. This produces a relatively complex space, since the way in which it will be divided by participants is not immediately obvious. This was clear from the results: participants in one chain structured the language by rotation, while in the other chains the structure was non-rotational. In Perfors and Navarro (2012), the meaning space has two dimensions (colour and size) that varied continuously in the first condition. These do not fully alleviate the

problem because the number of dimensions remains low, the structure of the space remains pre-specified, and the number of possible meanings remains finite.

1.5. Proposed research

There is a disconnect between the complex conceptual spaces used by humans to understand the world and the simple meaning spaces used in current experiments. For example, our perception of certain quality dimensions may be non-linear, as is the case in the perception of colour (perceived as circular) or pitch (which has a logarithmic relationship with frequency). We propose to address this issue here by implementing an infinite meaning space. This is an important step to take because a crucial aspect of language is the property of 'discrete infinity' – its ability to use a discrete set of symbols to talk about an infinite set of meanings (Studdert-Kennedy, 2005).

1.5.1. Research questions

This project attempts to answer two research questions. Firstly, does categorical structure arise from iterated learning when the meaning space is infinite in magnitude? Secondly, and if so, what is the effect on the structure of the signals? Do we observe the emergence of compositional structure such that recombinable linguistic units systematically map to specific areas of the meaning space? These questions will be evaluated using experimental iterated learning.

1.5.2. Hypotheses

Three outcomes are hypothesized. **Hypothesis I:** we expect the languages to become increasingly learnable over the course of the cultural generations. **Hypothesis II:** we expect to see the emergence of categorical and/or compositional structure as mechanisms for circumventing the bottleneck on transmission. **Hypothesis III:** given that Hypothesis I and Hypothesis II are supported, we expect to show that an increase in learnability can be explained by an increase in structure. The following chapter describes the methods that will be used to evaluate these hypotheses.

2. Methods

This chapter outlines the methods used to implement an infinite meaning space. Section 2.1 describes the experimental setup, including the procedure and stimuli. Section 2.2 describes the methods used to analyse the data, including a measure of learnability and structure. Section 2.3 describes the statistical methods used to evaluate the hypotheses.

2.1. Experimental setup

Since infinite meaning spaces have gone unexplored in experimental iterated learning, we have adopted a simple approach in this project using a procedure closely related to Kirby et al. (2008). The population model is a linear diffusion chain of 10 participants, each representing one cultural generation. Four chains were run in each of two experiments. Like Kirby et al. (2008), experiment 1 does not enforce expressivity, while experiment 2 does.

2.1.1. Recruitment of participants

Participants were recruited using the University of Edinburgh's SAGE website (Appendix A). It was described as a language learning task in which participants had to learn the language of the Flatlanders[4]. Forty participants completed experiment 1, and forty participants completed experiment 2. The experiment was conducted in accordance with the ethics procedures of the School of Philosophy, Psychology and Language Sciences. Each participant was paid £5.50 for completing the study. A £20 voucher redeemable at Amazon.co.uk was offered as a prize for the participant who was best able to learn the language[5].

2.1.2. Procedure

In the training phase, participants were exposed to three passes over a set of 48 stimuli for a total of 144 presentations. The stimuli were taken from an infinite meaning space of

[4] In reference to E. A. Abbott's (1884) novella, *Flatland: A romance of many dimensions*.

[5] In fact, the prize winner was drawn at random, since there is no objective way to determine the best learner. This prize was offered to incentivize participants to learn the language as well as possible.

Figure 2.1 Photograph of the experimental setup.

triangles and were presented with their labels in the Flatlander language. Triangle stimuli were presented first with their associated labels appearing after a 1-second delay. Participants also heard the label pronounced by a speech synthesizer simultaneously with the presentation of the written form. After every third triangle, the participant was shown one of the previous three triangles again and prompted to type its name. We refer to this as a mini test. Feedback was given in the form of a green[6] checkmark or red cross along with an audible sound. Over the course of the three passes, each of the 48 items was mini-tested exactly once. This training procedure was chosen to maintain participants' attention in what could otherwise be a very passive component of the experiment. In the test phase, participants saw 96 stimuli, none of which they had seen in training, and were prompted to type in the associated string for each one. No feedback was provided during the test. The task was explained to participants in a written brief (Appendix B) and orally (Appendix C). Figure 2.1 shows a photograph of the setup.

The experiment was coded in HTML and JavaScript with a PHP backend running locally on the computer terminal (Appendix F) and presented in Google Chrome (in full-screen "Presentation Mode") on a 13" MacBook Air. Figure 2.2 shows screenshots of the interface. The screen was positioned 70cm from the edge of the desk at an angle of 105°.

[6] The colour blue was also implemented as an alternative for red-green colourblind participants, but no male participant identified as colourblind.

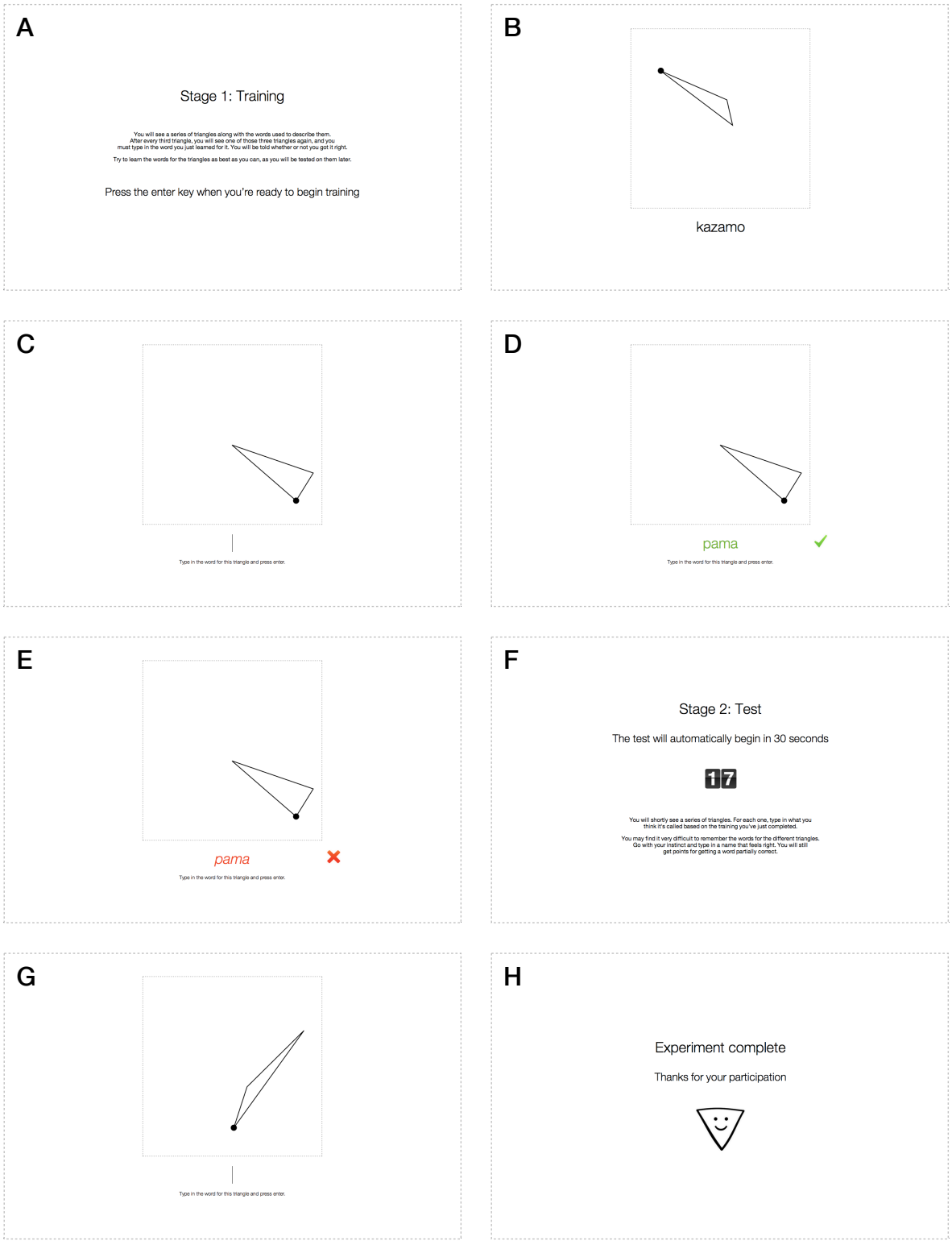


Figure 2.2 Screenshots of the computer interface. **(A)** Training instructions; the participant presses the enter key to begin. **(B)** Training item with its associated label. **(C)** After every third training item, the participant is shown one of the previous three training items and prompted to type its name. **(D)** If correct, the participant sees a green checkmark and the word is highlighted in green. **(E)** If incorrect, the participant sees a red cross and the correct answer is shown in red italics. Screenshots **(B)** through **(E)** are then repeated until the entire training set has been covered three times with each of the training items being mini-tested exactly once. **(F)** Thirty-second break between the training and test phases; includes reminder instructions for the test. **(G)** Test item: the participant sees a novel stimulus and is prompted to enter its name. This is repeated 96 times until both the DYNAMIC and STABLE SETS have been covered in an interleaved order. **(H)** Completion message.

An Apple Wireless Keyboard was used for input, and a custom keyboard layout blocked all keys except the lower-case alphabetical keys, enter key, and backspace key. Vocalizations of the strings were played through Sennheiser HD 219 headphones. On finishing the experiment, participants completed a post-test questionnaire (Appendix D) and were offered a debrief to take away with them (Appendix E).

2.1.3. Transmission

Figure 2.3 illustrates how the languages were transmitted. A set of 48 stimuli (referred to as a DYNAMIC SET) is randomly generated at every generation, such that no two participants will ever be trained on the exact same stimuli. The purpose of the DYNAMIC SET is to model the infinity of nature, where no two people ever observe the exact same meaning. In addition, there is a single set of 48 stimuli, the STABLE SET, that remains the same for all participants in both experiments. Participants are never trained on the STABLE SET, only tested on it. Its purpose is to provide a constant set of stimuli on which we can measure learnability and structure.

The training phase involves learning the mapping between the previous participant's output WORD SET and DYNAMIC SET (presented in randomized order). The test phase involves providing labels for (a) the 48 stimuli in a randomly generated DYNAMIC SET and (b) the 48 stimuli in the STABLE SET (presented in randomized order) for a total of 96 test items. These two components of the test phase are interleaved to avoid the possibility that participants might be fatigued by the time they reach the second component. Each participant returns two WORD SETS as output, a WORD SET that labels their DYNAMIC SET (for transmission to the next participant), and a WORD SET that labels the STABLE SET (for analysis). WORD SETS referred to as 0 and 0′ in Figure 2.3 are randomly generated sets of 48 words used to initialize a chain.

2.1.4. Triangle stimuli

The STABLE SET was randomly generated prior to running the experiments; the triangle stimuli in DYNAMIC SET 0 were randomly generated prior to beginning a chain (`initial_set_generator.py`; Appendix F). The triangles in subsequent DYNAMIC SETS were randomly generated on demand during each participant's test phase and were saved at the end of the test as a set of coordinates for reconstruction during the subsequent participant's training. To generate a triangle stimulus, three points are chosen at random in a 480×480-pixel space and joined together with 2-pixel-wide lines. The space was enclosed in a

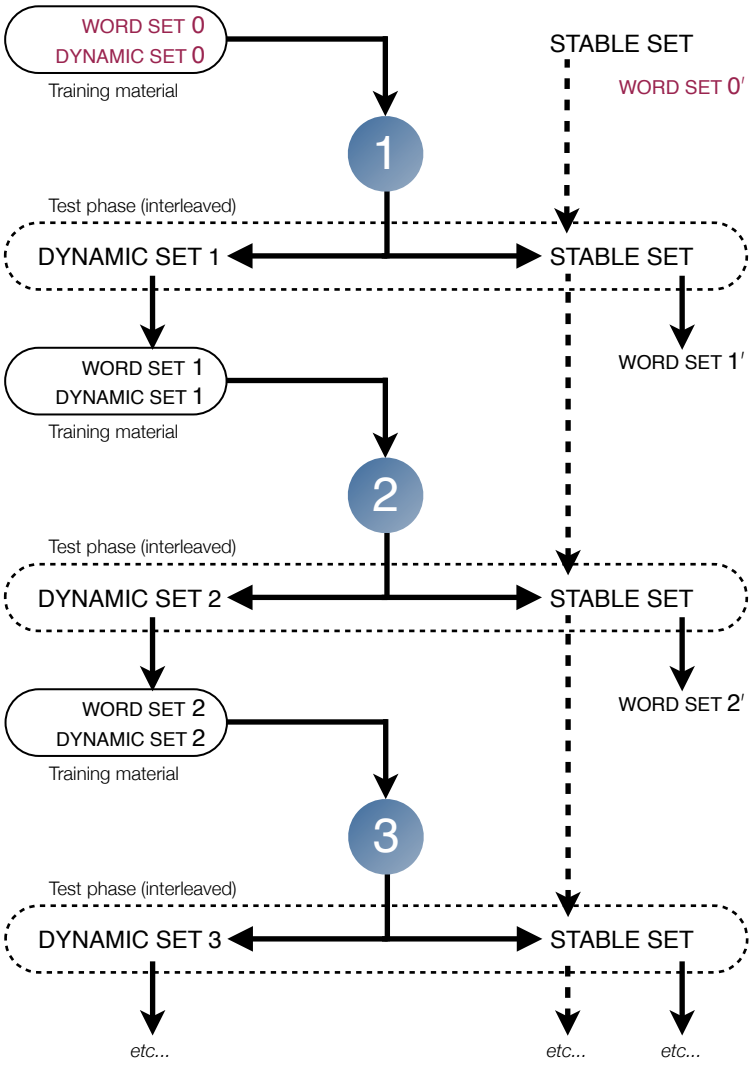


Figure 2.3 Diagram illustrating the experimental procedure. The blue circles represent the participants in a diffusion chain. Each participant is trained on one set of experimental stimuli, but tested on two different sets: a randomly generated set (referred to as a DYNAMIC SET) and a set that remains the same across both experiments (referred to as the STABLE SET). A DYNAMIC SET, along with the words used to label it, forms the input for the subsequent participant. The sets presented in red are randomly generated to initialize each chain.

500×500-pixel dashed box. One point (determined randomly) has a black circle with a radius of 8 pixels placed over it, and is referred to as the orienting spot. Its function is to give the participant some context about which way the triangle is oriented. The number of distinct stimuli that can be generated by this simple algorithm is approximately 6×10^{15} (6 million billion)[7]. To all intents and purposes this represents an infinite number of possible triangles, limited only by the density of the pixels on the display[8]. See Figure 2.4 for two

[7] $(480^2)^3/2$. Divide by 3 if assuming the orienting spot does not distinguish a vertex.

[8] If we assume that participants cannot distinguish between two triangles whose vertices are within r pixels of each other, the probability of having two triangles perceived as identical can be approximated using $(480^2/\pi r^2)^3/2$. At 10 pixels, this probability is approximately 1 in 200 million. At 20 pixels it falls to 1 in 3 million, which is still hundreds of times greater than the total number of triangles observed by all 80 participants.

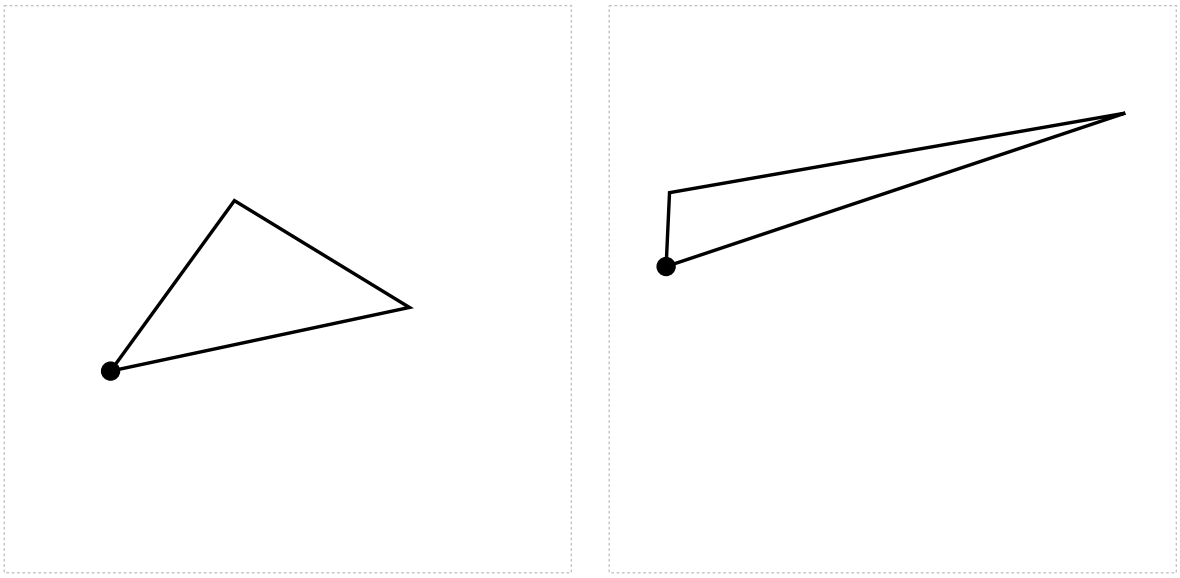


Figure 2.4 Two examples of the triangle stimuli in the STABLE SET. Lines thickened for clarity.

examples of the triangle stimuli and Appendix G for the full STABLE SET.

2.1.5. Linguistic stimuli

The initial WORD SETS were generated randomly from a finite set of syllables. A syllable consists of a consonant from the set {d, f, k, m, p, z} concatenated with a vowel from the set {a, i, o, u} for a total of 24 possible syllables. The consonant set was chosen because it covers a range of places and manners of articulation, as well as both voiced and unvoiced consonants, and the letters are clear and unambiguous for a native English speaker. The vowel set, pronounced [ɑː], [iː], [əʊ], [uː] respectively, was chosen because it represents a distinct set of sounds in the vowel space. Strings were generated by simply concatenating between 2 and 4 syllables at random (`language_generator.py`; Appendix F).

In previous work (Carr, 2012), I used a speech synthesizer to vocalize the strings during the training phase. There are at least four advantages to this approach. Firstly, the combination of the written and spoken modalities provides increased stimulation for participants. Secondly, if the participant is momentarily distracted or looks away from the screen, they will at least hear the word. Thirdly, the use of vocalizations provides participants with a systematic phonological system, so they do not need to consider how to pronounce or sub-vocalize a word. Fourthly, all participants hear the words pronounced under the same systematic phonological system, reducing the chance that two participants might analyse a word differently because they happened to pronounce or sub-vocalize it differently. To produce a synthesized vocalization, we convert the string into a sequence of

machine-readable phonemes and use the MacinTalk speech synthesizer (Apple Computer, 2006) to output an MPEG4 audio file containing the rendered word. For example, the word *pokifu* would be transformed into machine readable `pOWk1IYfUW` and then rendered in audio as [pəʊˈkiːfuː]. Vocalizations were automatically generated using a Python script (`vocalize.py`; Appendix F)[9].

2.1.6. Experiment 2 modification

The experiment described above does not enforce expressivity, and it is likely that the languages will rapidly become underspecified (see Kirby et al., 2008). This is not a problem for observing the emergence of categorical structure, since even a handful of words can carve up the meaning space in a systematic way. However, to test whether we would see the emergence of compositional structure, the procedure was modified in experiment 2 so that participants could not use the same string more than three times[10] to label items in their DYNAMIC SET. Upon attempting to enter a string that had already been used three times, the participant was presented with the message 'Ooops! You've used this word too often. Please use another word'. The participant was free to use the same string as often as they wanted to describe triangles in the STABLE SET[11]. Experiment 2 is identical to experiment 1 in all other regards. This procedure was inspired by a similar method used in Verhoef (2012) in which participants were forced to use distinct songs to prevent underspecification. The enforcement of expressivity parallels the original iterated learning experiments of Kirby et al. (2008).

2.2. Data analysis

This section describes the two measures used to analyse the data: a measure of learnability

---

[9] Before a participant began the experiment, these audio files were manually checked in case of error, inconsistency, or disfluency in the rendering. Occasionally, modifications would be made to the transformation rules in the Python script to account for unusual combinations of vowels or consonant clusters that emerged, which would then be pronounced in a single consistent way through to the end of the chain.

[10] This parameter, which we refer to as λ (for limit), was set based on results from experiment 1, which suggested that $\lambda = 3$ would be flexible enough to allow for compositional languages but limiting enough to prevent runaway underspecification. If λ is set too high, the participant will rarely be asked to use a different word resulting in underspecification to a holistic system. Conversely, a low λ restricts the type of language that can evolve to one that uses a large number of distinct words, which may not be conducive to a natural way of structuring the meaning space nor to the emergence of compositional structure. A low λ also forces participants to frequently change the words they want to use, which may be unfair.

[11] This is because the STABLE SET is not passed on to the next generation, so duplicates in this set will not lead to runaway underspecification. By opening up the STABLE SET to unlimited duplication, the participant is only constrained by the λ parameter where it really matters for preventing underspecification.

and a measure of structure.

2.2.1. Measuring learnability

Transmission error is used as a proxy for learnability under the assumption that greater error in predicting the words that the previous participant applied to items in the STABLE SET implies the presence of a less learnable language (and vice versa). Transmission error is quantified by taking the mean normalized Levenshtein edit distance[12] (Levenshtein, 1966) between the strings used to describe items in the STABLE SET at generation i and the corresponding strings at generation $i-1$. More formally, transmission error E for generation i is given by

$$E(i) = \frac{1}{|M|} \sum_{m \in M} \frac{\mathrm{LD}(s_i^m, s_{i-1}^m)}{\max(\mathrm{len}(s_i^m), \mathrm{len}(s_{i-1}^m))}, \tag{2.1}$$

where the Levenshtein edit distance (LD) between string s for meaning m at generation i and the corresponding string in the preceding generation is normalized by dividing by the length of the longer string. This measure of error is expressed over the interval [0,1], where 0 represents identity and 1 represents a lack of common characters. The mean edit distance for all $m \in M$ gives the final measure of error.

2.2.2. Measuring structure

Our languages are essentially mappings between signals and meanings. To measure how structured these mappings are, we simply correlate the dissimilarity between pairs of strings with the dissimilarity between pairs of triangles for all $n(n-1)/2$ pairs. Standard parametric statistics are not suitable for this, since the pairwise distances are not independent from each other (see Cornish, 2011, p. 92–93 for some discussion). To get around this problem, we perform a Mantel test (Mantel, 1967) which compares the correlation for the veridical string-meaning alignment against a distribution of correlations for 50,000 Monte-Carlo permutations of the signal-meaning pairs[13]. This yields a standard score (z-score) quantifying the significance of the veridical correlation. The normalized Levenshtein distance was used to measure the dissimilarity between pairs of strings. The

[12] The minimum number of deletions, insertions, and substitutions required to transform one string into another.

[13] As the number of simulations is increased, the scores approach their true value for all 48! permutations of the signal-meaning pairs. 50,000 simulations is therefore a tradeoff, being accurate to approximately 1 or 2 decimal places, but taking around 20 hours to compute for all metrics for all 80 participants.

following section describes the the more complex problem of measuring the dissimilarity between pairs of triangles.

2.2.3. Triangle dissimilarity metric

The distance between two triangles is defined as the sum of the Euclidean distances between corresponding vertices. The orienting spot on one triangle automatically corresponds to the orienting spot on the other (i.e. vertex A_1 corresponds to vertex B_1). To determine whether A_2 should correspond to B_2 or B_3, and consequently whether A_3 should correspond to B_2 or B_3, we take the correspondence that yields the smaller sum of Euclidean distances. Thus, the distance function between two triangles A and B, referred to as d_T, is given by

$$d_T(A,B) = d_E(A_1,B_1) + \min[d_E(A_2,B_2) + d_E(A_3,B_3), d_E(A_2,B_3) + d_E(A_3,B_2)], \qquad (2.2)$$

where $d_E(A_i,B_j)$ is the Euclidean distance between points A_i and B_j, which itself is given by $\sqrt{(x_{A_i} - x_{B_j})^2 + (y_{A_i} - y_{B_j})^2}$. For all triangles A, B, and C in a metric space X, d_T satisfies the four conditions on a distance metric: non-negativity, i.e. $d_T(A,B) \geq 0$; the coincidence axiom, i.e. $d_T(A,B) = 0$ iff $A = B$; symmetry, i.e. $d_T(A,B) = d_T(B,A)$; and the triangle inequality, i.e. $d_T(A,B) \leq d_T(A,C) + d_T(C,B)$.

While our triangle distance function d_T provides a first approximation of the distance between two triangles, it is deficient in a number of ways. For example, it does not consider how location, orientation, and size (nor combinations of these properties) affect the perceived similarity between two triangles. One simple way to circumvent this problem is to measure d_T between triangle A and a transformation of triangle B that brings it into closer alignment with A. This allows us to eliminate the effect of one or more transformations at a time and observe how it affects d_T. The transformations that we consider here are translation, rotation, scaling, and all combinations of these three. This gives us eight distance metrics, as listed in Table 2.1, each of which considers and ignores different geometrical properties[14]. The metrics are said to measure d_T 'up to' (i.e. disregarding) the transformations used to bring the triangles into alignment.

---

[14] Since we have eight distance metrics, we compute eight separate structure scores, each score reflecting different combinations of geometrical properties. This is necessary because we have no way of determining *a-priori* the kinds of properties that are likely to be salient, nor how the salience of properties will be weighted. The lack of a single structure score could be seen as a disadvantage; however, this approach actually offers a few advantages: (1) it allows us to identify which properties participants find salient; (2) it helps us to detect lineage specificity; and (3) it makes fewer assumptions about what a structured language should look like.

Table 2.1 List of triangle distance metrics and the geometrical properties that they ignore and consider

Distance metric	Properties ignored	Properties considered
d_T	—	shape, location, orientation, size
d_T up to translation	location	shape, orientation, size
d_T up to rotation	orientation	shape, location, size
d_T up to scale	size	shape, location, orientation
d_T up to rigid motion	location, orientation	shape, size
d_T up to scaled translation	location, size	shape, orientation
d_T up to scaled rotation	orientation, size	shape, location
d_T up to scaled rigid motion	location, orientation, size	shape

***d_T* up to translation** (d_{T_t}) is defined as d_T between triangle A and a translation of triangle B, denoted B', that brings its centroid (B_c) into alignment with the centroid of triangle A (A_c). Thus, each new vertex B'_i is given by $B_i + (A_c - B_c)$. ***d_T* up to rotation** (d_{T_r}) is defined as d_T between triangles A and B after they have been rotated to point directly upwards. The angle of rotation θ for any triangle A is the angle of the line passing through its centroid and orienting spot, and is given by

$$\theta(A) = \begin{cases} \arccos \frac{p^2+q^2-r^2}{2pq} & \text{if } x_{A_1} \leq x_{A_c} \\ 0 - \arccos \frac{p^2+q^2-r^2}{2pq} & \text{if } x_{A_1} > x_{A_c} \end{cases}, \tag{2.3}$$

where $p = d_E(A_1, A_c)$, $q = d_E(A_c, (x_{A_c}, 0))$, and $r = d_E((x_{A_c}, 0), A_1)$. To rotate A into an upright orientation, the triangle is first translated so that its centroid lies at the origin, then each new x-coordinate is calculated according to $x \cos\theta - y \sin\theta$, and each new y-coordinate is calculated according to $x \sin\theta + y \cos\theta$, before translating the triangle back to its original position. ***d_T* up to scale** (d_{T_s}) is defined as d_T between a scaling of triangles A and B so that both have a perimeter of 750 pixels[15]. It is necessary that both triangles are scaled to the same arbitrary size, as opposed to scaling one triangle to the size of the other, so that d_{T_s} meets the condition of symmetry, i.e. $d_{T_s}(A, B) = d_{T_s}(B, A)$ [16]. To scale triangle A so that its perimeter $P = 750$, we determine the scaling factor $f = 750/P(A)$, translate the triangle so

[15] This number was chosen because it results in scaled triangles that are approximately representative of the average perimeter of a triangle in our 480×480-pixel space, which was determined to be ≈751 pixels based on the mean perimeter of 1 million randomly generated triangles. Perimeter was used rather than area because scaling very thin, line-like triangles based on area produces extremely long triangles that dramatically skew the correlation with string-edit distance.

[16] To see why, imagine that A is a large triangle and B is a small triangle. Scaling B to the area of A results in two large triangles; scaling A to the area of B results in two small triangles. The Euclidean distances between vertices will not be equal (or even proportional) under these two scalings, resulting in an asymmetrical metric.

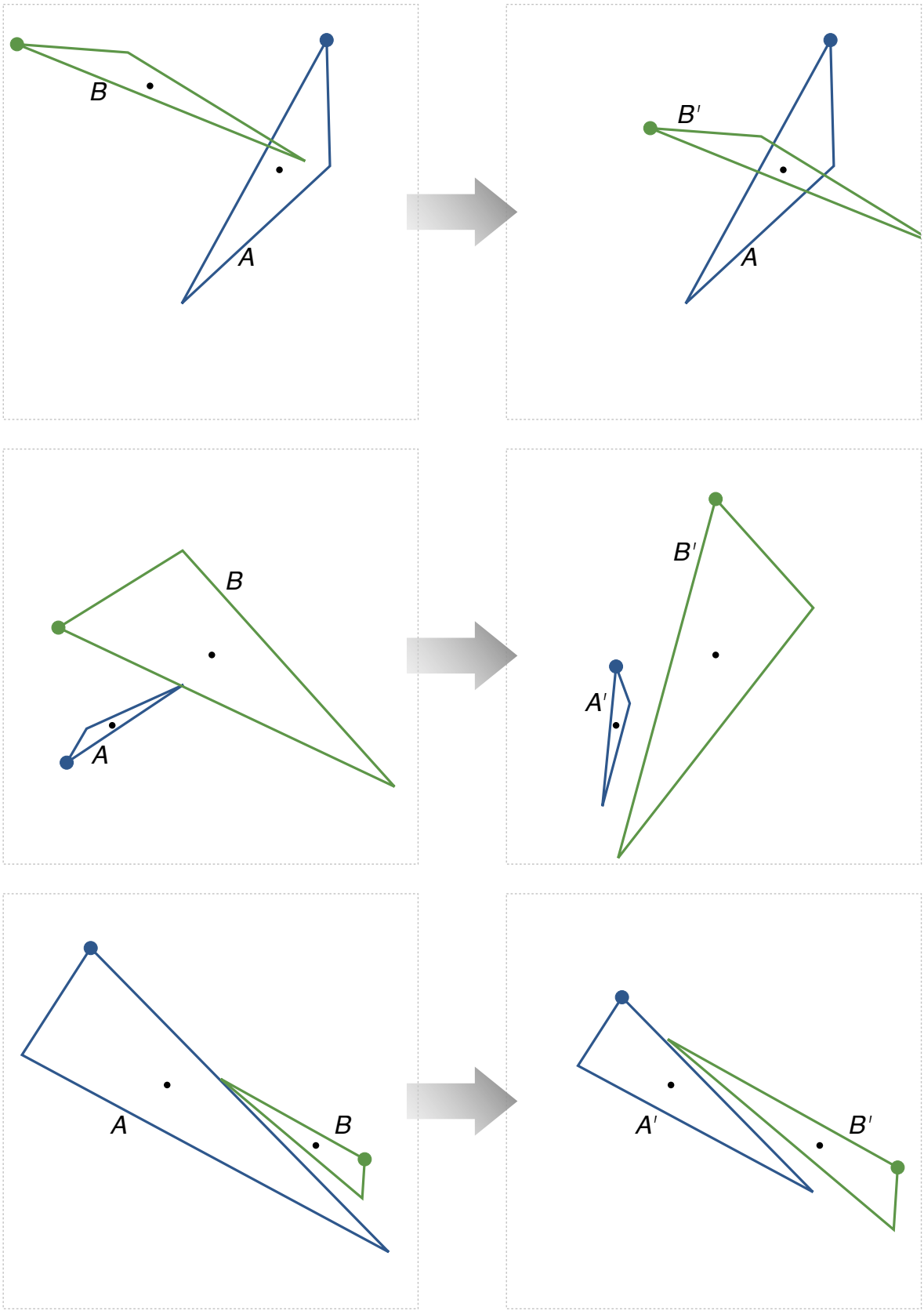


Figure 2.5 (Top) Translation of triangle *B* so that its centroid aligns with that of triangle *A*. **(Middle)** Rotation of triangle *A* and triangle *B* around their centroids so that they both point upwards. **(Bottom)** Scaling of triangle *A* and triangle *B* around their centroids so that they have equal perimeter. Centroids are marked by black dots.

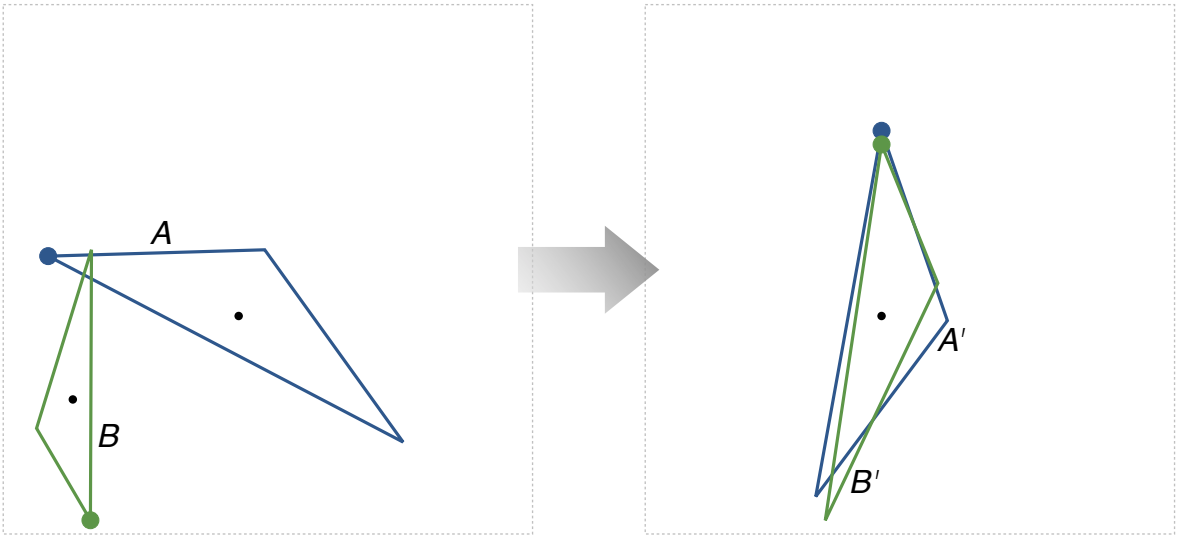


Figure 2.6 Illustration of a scaled rigid motion transformation. Both triangles are rotated to an upright orientation and scaled to equal perimeter. Triangle *B* is also translated so that its centroid aligns with that of triangle *A*. In this way, the distance between *A*′ and *B*′ quantifies the difference in shape only, ignoring location, orientation, and size.

that its centroid lies at the origin, and then multiply each x- and y-coordinate by f, before translating the triangle back to its original position. The translation, rotation, and scaling transformations are illustrated in Figure 2.5.

We now turn to the composite transformations. $\boldsymbol{d_T}$ **up to rigid motion** ($d_{T_{rm}}$) is simply d_T between triangle A and a translation and rotation (i.e. a rigid motion) of B following the procedures described above. Likewise, $\boldsymbol{d_T}$ **up to scaled translation** ($d_{T_{st}}$) is d_T between a scaling of A and a translation and scaling of B. $\boldsymbol{d_T}$ **up to scaled rotation** ($d_{T_{sr}}$) is d_T between a rotation and scaling of A and a rotation and scaling of B. Finally, $\boldsymbol{d_T}$ **up to scaled rigid motion** ($d_{T_{srm}}$) is d_T between a scaling and rotation of A and a scaling and rigid motion of B. The scaled rigid motion transformation is illustrated in Figure 2.6. The code for all geometrical transformations and measurements described above can be found in `geometry.py` (Appendix F).

2.3. Statistical methods

The measure of structure discussed in the previous section has a statistical test of significance built in which informs us about how unlikely it is that a given correlation between string dissimilarity and meaning dissimilarity could have occurred by chance. However, we would also like to test the hypothesis that structure, as well as learnability, increase with generation number. Recent iterated learning experiments have used Page's

trend test (Page, 1963) to evaluate the significance of the cultural evolutionary trend (e.g. Smith & Wonnacott, 2010; Verhoef, 2012). The test evaluates the alternative hypothesis that generation 1 < generation 2 < ... < generation *n* against the null hypothesis that generation 1 = generation 2 = ... = generation *n* (i.e. a repeated measures and ordered counterpart to the Spearman rank correlation coefficient). Page's *L* will be the primary statistic for evaluating Hypothesis I (learnability increases) and Hypothesis II (structure increases). A Python implementation of Page's trend test can be found in `page.py`, (Appendix F).

To evaluate Hypothesis III, we can simply correlate structure scores for generation *i* with learnability scores for generation *i*+1. The idea here is that, if a participant produces a more structured language, the participant in the following cultural generation should find that language easier to learn (and vice versa)[17]. This intergenerational approach is more powerful than the typical use of a correlation, since it is possible to observe the direction of causation: a positive correlation, for example, implies that structure leads to languages that are more learnable in the subsequent generation (because it cannot be the case that more learnable languages cause structure to emerge in the previous generation). Of course, the normal caveat that correlation does not imply causation still applies.

[17] A solution to the problem of non-independence of data points is discussed in Section 3.3.3.

3. Results

This chapter provides the results from the experiments. Section 3.1 describes the results for experiment 1. Section 3.2 describes the results for experiment 2. Section 3.3 offers a comparison of the experiments.

3.1. Experiment 1

Forty participants (20 female) completed experiment 1 during the period 1st—12th July 2013. The mean age was 23.1 years ($SD = 3.15$). Since generation number is our independent variable, it should not be correlated with factors incidental to the experiment. There was no significant correlation between generation number and the time of the day ($r = 0.167$, $n = 40$, $p = 0.303$) nor the day of the week ($r = 0.218$, $n = 40$, $p = 0.177$) that the experiment was completed on. Generation number was also not correlated with sex ($r = 0.113$, $n = 40$, $p = 0.486$) or language experience ($r = 0.012$, $n = 40$, $p = 0.94$). However, it was significantly correlated with age ($r = 0.341$, $n = 40$, $p = 0.031$). The average age in the first five generations was 22.1 years, while the average age in the latter five generations was 24.1 years[18].

3.1.1. Number of unique strings

The number of unique strings in the languages rapidly collapsed down to an average of 5 in each of the DYNAMIC and STABLE SETS by generation 10. These results are shown in Figure 3.1. This falling trend is highly significant both for the DYNAMIC SET ($L = 1990$, $m = 4$, $n = 11$, $p < 0.001$) and the STABLE SET ($L = 1993$, $m = 4$, $n = 11$, $p < 0.001$).

3.1.2. Learnability

The results for transmission error are shown in Figure 3.2.A, which shows that error tends to fall as subsequent participants are increasingly able to predict the words that the previous participant applied to items in the STABLE SET. This falling trend is highly significant ($L = 1514$, $m = 4$, $n = 10$, $p < 0.001$). However, given that the number of

[18] We did not investigate this factor further, since it seemed unlikely that a difference of two years would have any significant impact on the validity of the experiment.

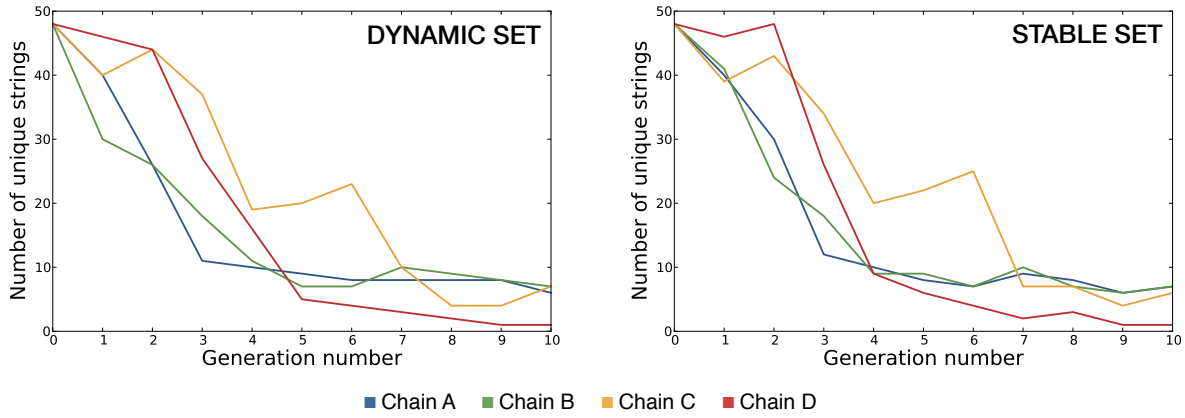


Figure 3.1 The number of unique strings over the course of 10 generations in chains A–D for the dynamic set (left) and stable set (right). The number of unique strings rapidly decreases from 48 down to an average of 5.

unique strings in the languages tends to decrease with generation number, it is unsurprising that transmission error also decreases, since chance guesses are increasingly more likely to be correct. To account for this, we computed a distribution of transmission error scores for 100,000 Monte-Carlo permutations of each participant's STABLE SET and calculated a *z*-score for the veridical score as compared to the Monte-Carlo sample. This transformation quantifies how unlikely it is that a given error score could have been generated by chance (or in other words the non-randomness of the alignment between consecutive STABLE SETS) and is therefore a better estimate of learnability[19]. The results from this transformation are shown in Figure 3.2.B. This graph suggests that learnability tends to increase over the ten generations. One data point (D10) is undefined under this transformation because the

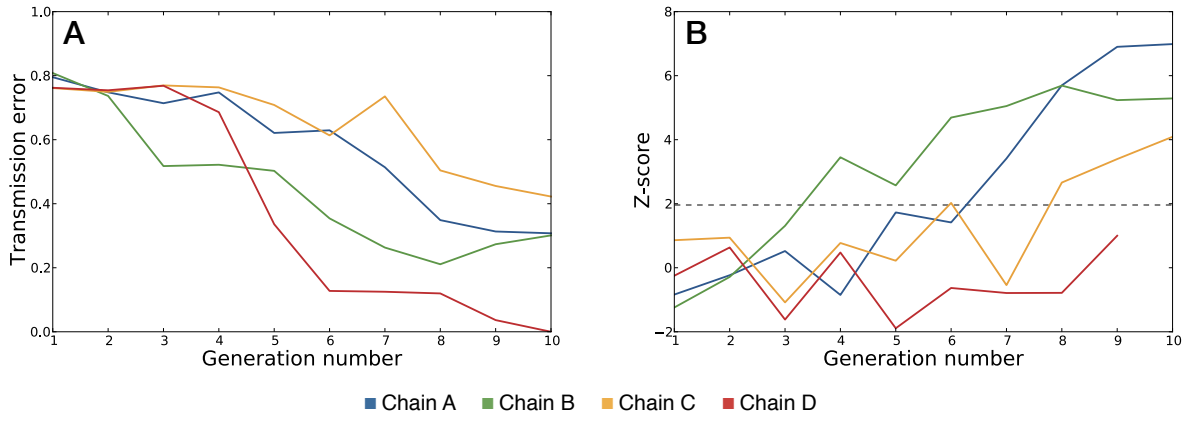


Figure 3.2 (A) Transmission error over the course of 10 generations for chains A—D. **(B)** *Z*-score transformation of the error scores by comparing the veridical score to a distribution of error scores for Monte-Carlo permutations of the STABLE WORD SET. The data point for D10 is undefined, since all triangles were referred to by a single word at this generation. The dashed line in (B) show the two-tailed 95% confidence level.

[19] This procedure is novel to the present project and should be applied in future research of this kind, especially where there is a dramatic collapse in the number of words.

participant used a single word to refer to all triangles. Since Page's trend test cannot handle missing data points, the test was run over the first nine generations only and was highly significant for a rising trend ($L = 1038$, $m = 4$, $n = 9$, $p < 0.001$). The importance of this z-score transformation is made particularly clear by chain D. The results for transmission error suggest that participants in this chain made the least amount of error, but, on consulting the z-score graph, it is clear that these levels of error could easily have been generated by chance.

Besides looking at transmission error, there are some other ways we can show that the languages are becoming easier to learn. Firstly, we can look at the amount of error participants made during mini tests. This decreased significantly from an average of 29.3% error in generation 1 to 6.5% error in generation 10 ($L = 1442.5$, $m = 4$, $n = 10$, $p < 0.001$). Secondly, we can look at how quickly participants responded to the test items, with quicker response times indicating a more accessible language. The time spent on each test item decreased significantly over the course of the experiment from 8.23 seconds in generation 1 to 3.42 seconds in generation 10 ($L = 1425$, $m = 4$, $n = 10$, $p < 0.001$). Finally, participants were asked in the post-test questionnaire to rate how difficult it was 'to learn and recall the words'. Participants' ratings decreased significantly over the ten generations from an average of 9.5 in generation 1 to 5.75 in generation 10 ($L = 1381$, $m = 4$, $n = 10$, $p < 0.001$)[20].

3.1.3. Structure

The measure of structure for each of the eight triangle distance metrics is given in Figure 3.3. In chain D, which collapsed to a small number of words, this approach performs poorly and the scores are therefore only provided up to generation 6, after which the number of unique strings in the STABLE SET drops to 2. The Bonferroni correction has been applied to adjust for the fact that eight separate metrics have been used to quantify the structure in the languages, yielding a two-tailed 95% confidence level for significantly structured languages of $z = 2.734$. Table 3.1 gives the mean and standard deviation of the z-scores for each metric.

[20] Of course, these three measures are also subject to the same criticism we levelled at our transmission error measure: it is unsurprising that they decrease given the collapse in the number of unique strings.

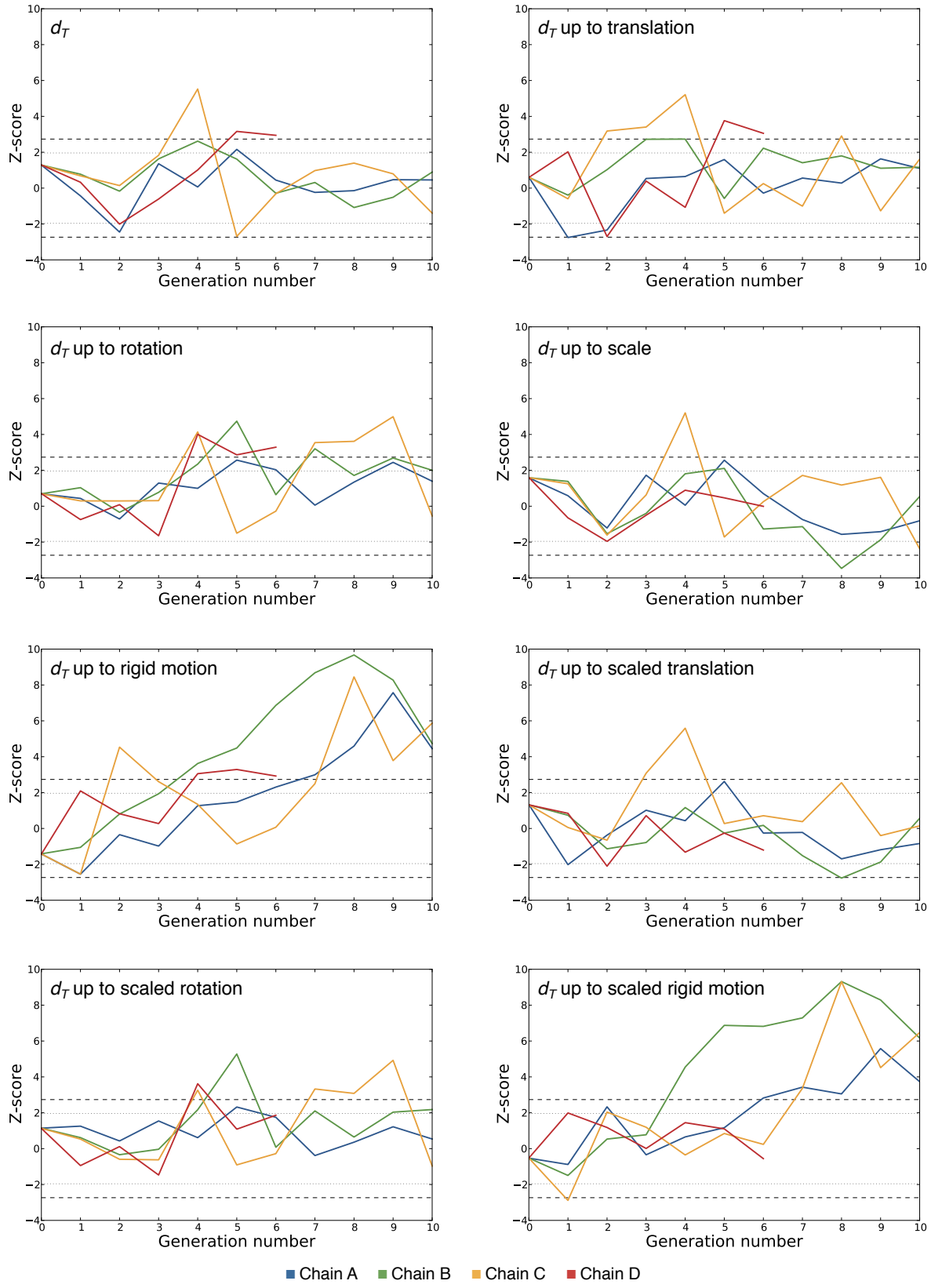


Figure 3.3 Plots showing the structure in the languages under the eight triangle distance metrics. Note that the scores for participants D7 through D10 are undefined here, since these participants used a very small number of words to describe the triangles, making this approach to measuring structure unsuitable. The dotted lines show the *z*-scores equivalent to $\alpha = 0.05$ and the dashed lines show the *z*-scores equivalent to $\alpha = 0.00625$ which corrects for eight multiple comparisons using the Bonferroni correction. Scores that lie outside of the dashed lines are significantly unlikely to have occurred by chance alone.

Table 3.1 Mean and standard deviation of the structure *z*-scores for each distance metric

Distance metric	Mean	SD
d_T	0.61	1.514
d_T up to translation	0.86	1.753
d_T up to rotation	1.403	1.662
d_T up to scale	0.169	1.659
d_T up to rigid motion	2.532	3.231
d_T up to scaled translation	0.137	1.571
d_T up to scaled rotation	1.123	1.546
d_T up to scaled rigid motion	2.461	3.119

These results suggest that the type of structure that is emerging in the languages is primarily concerned with size and shape, since the $d_{T_{rm}}$ and $d_{T_{srm}}$ metrics give by far the highest scores[21]. Furthermore, there is a highly significant upward trend for scores based on $d_{T_{rm}}$ ($L = 1461$, $m = 3$, $n = 11$, $p < 0.001$) and $d_{T_{srm}}$ ($L = 1470$, $m = 3$, $n = 11$, $p < 0.001$)[22]. Given these results, we focus primarily on size and shape in the following analyses, but note that the languages may be structured in other ways that our metrics fail to quantify.

Any triangle with perimeter P has area $\leq P^2/(12\sqrt{3})$ with equality iff the triangle is equilateral. We use this fact to approximate the shape property, since deviations from the maximum area for a given perimeter correspond to less equilateral triangles[23]. To visualize the relationship between size and shape, we produced scatterplots showing perimeter against area for the participant in each chain who had the strongest $d_{T_{rm}}$-based structure score (see Figure 3.4). Each point is colour-coded according to the word used to describe that triangle, allowing us to judge how the words correspond to the underlying shape-size conceptual space.

Participant B8 produced 7 unique strings, but two of these, *mamofudu* and *mamoziki*, appear to be typos on *mamofudo* and *mamozuki* respectively (this is especially clear when you look at where these two words lie in the space). With this is mind, there appear to be five words in the language at this generation. Figure 3.5 shows the triangles that these five words describe, along with a prototypical form of each triangle produced by averaging

[21] This bias is supported by the literature (e.g. Landau, Smith, & Jones, 1988).

[22] Chain D has been excluded in these runs of Page's trend test, since several of the data points are undefined. The tests remain significant even if the randomly generated generation-0 data points are excluded: $L = 1100$, $m = 3$, $n = 10$, $p < 0.001$ (for both d_{Trm} and d_{Tsrm}).

[23] Of course, this does not fully capture the concept of shape, but it is convenient for visualizing the relationship between size and shape in two-dimensional plots.

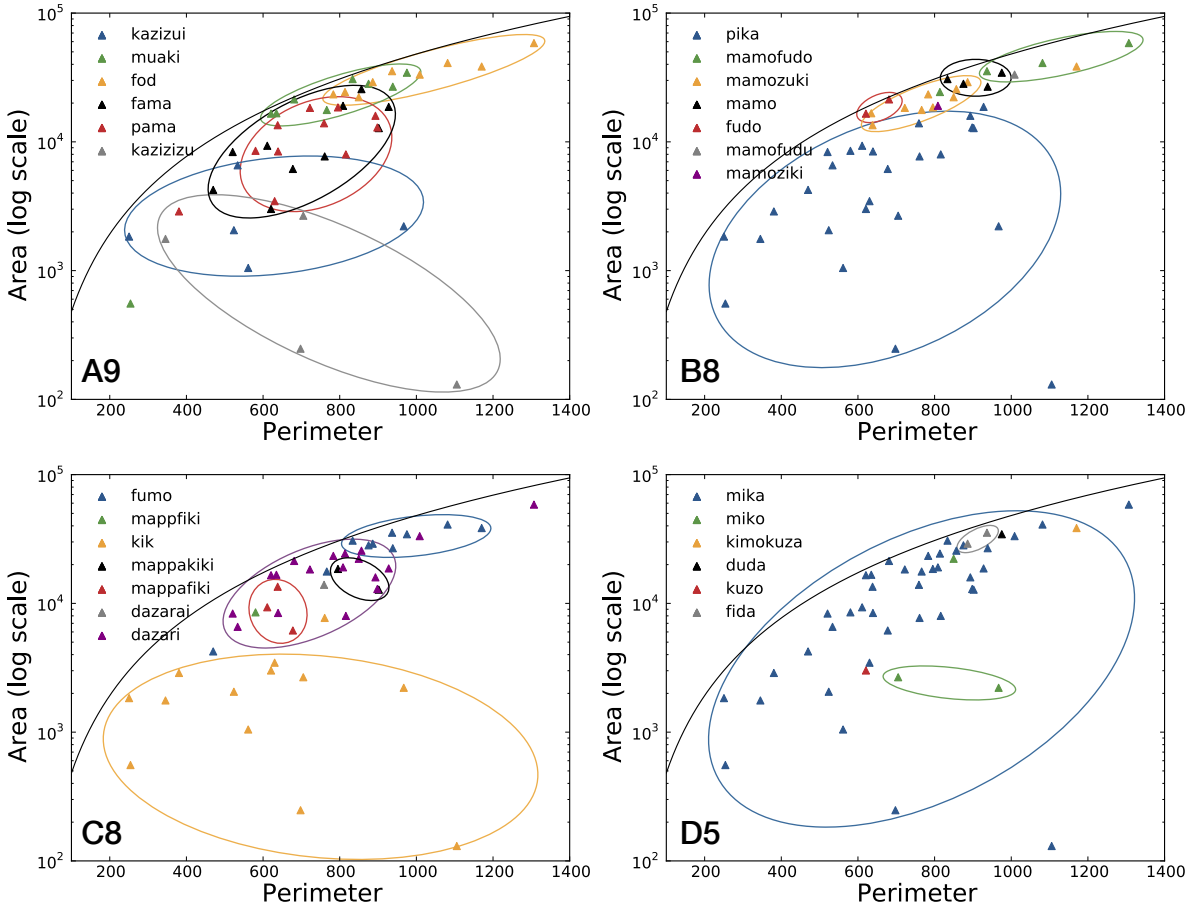


Figure 3.4 Scatterplots showing perimeter against area for all triangles in the STABLE SET for four participants. Each point corresponds to one triangle and is coloured according to the word used to describe it. The space inside the plots roughly corresponds to shape vs. size. Black curves show the relationship between area and perimeter for perfectly equilateral triangles. As you move along the curve, the triangles go from small to large; as you move away from the curve, the triangles become skinnier. Ellipses show interpretations of how the space may be carved up.

them together[24]. These graphics give us a sense of the central tendency (the prototype) and the variance (the triangles themselves) within a set of triangles described by a single word. *Pika* is used to describe small or thin triangles. *Mamofudo* is used for very large triangles. The distinction between the intermediate triangles is less clear, but one interpretation could be that *mamozuki* is used for right-angled triangles, *mamo* is used for isosceles triangles, and *fudo* is reserved for the most equilateral triangles.

The language in chain A at generation 9 divides the shape-size space a little differently. The words *kazizui* and *kazizizu* are used for small and particularly thin triangles. Large triangles are referred to by the word *fod*. Medium-sized triangles are referred to by two sets

[24] All triangles were translated to the centre of the space and then rotated upright with either the orienting spot or the vertex with the smallest angle on top; vertices 2 and 3 were then relabelled, if necessary, to ensure that the second vertex was always to the left of the first; finally, the coordinates were averaged together. This method is not foolproof, but it gives a rough idea of the prototypical form of a given set of triangles.

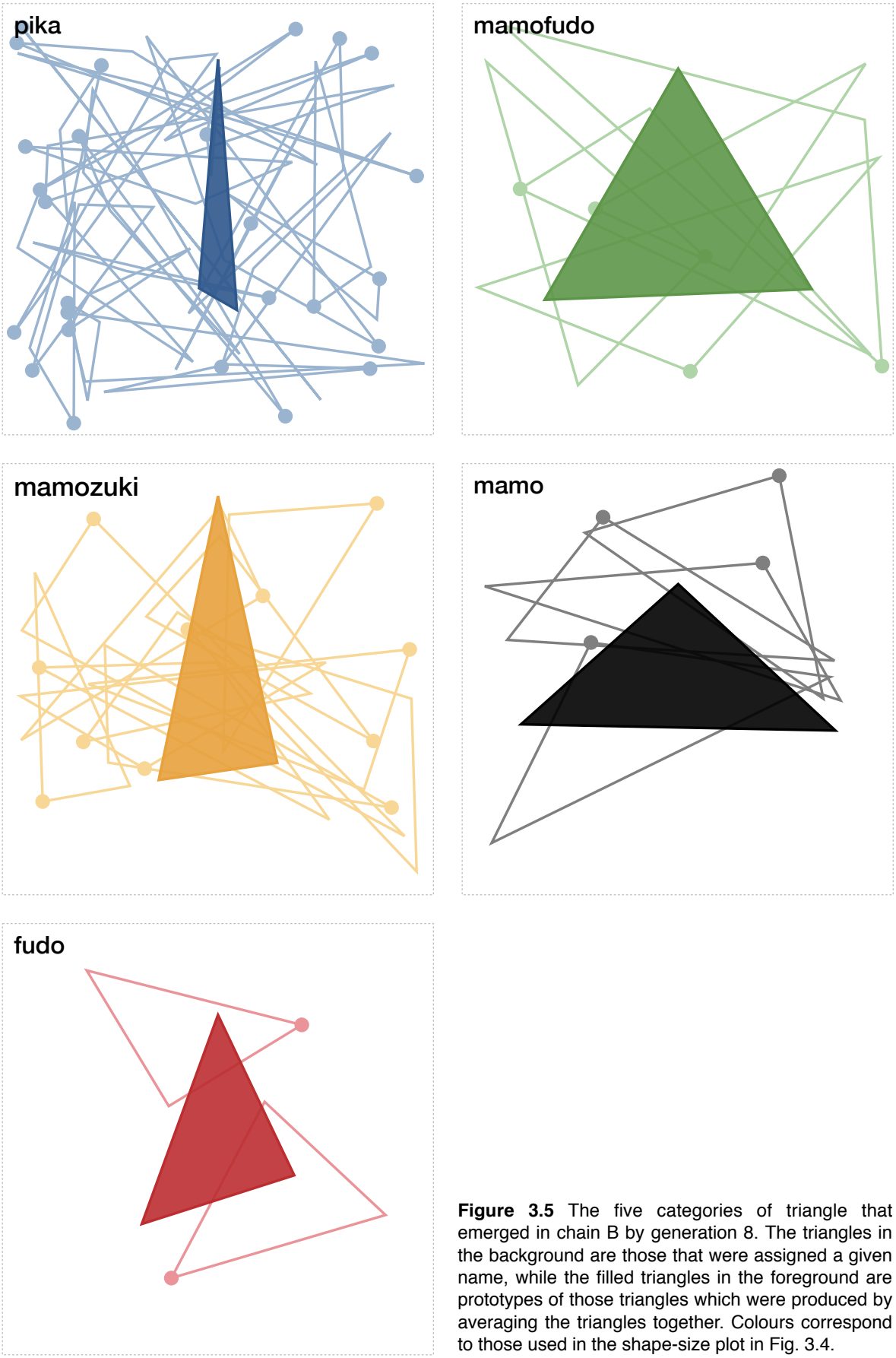


Figure 3.5 The five categories of triangle that emerged in chain B by generation 8. The triangles in the background are those that were assigned a given name, while the filled triangles in the foreground are prototypes of those triangles which were produced by averaging the triangles together. Colours correspond to those used in the shape-size plot in Fig. 3.4.

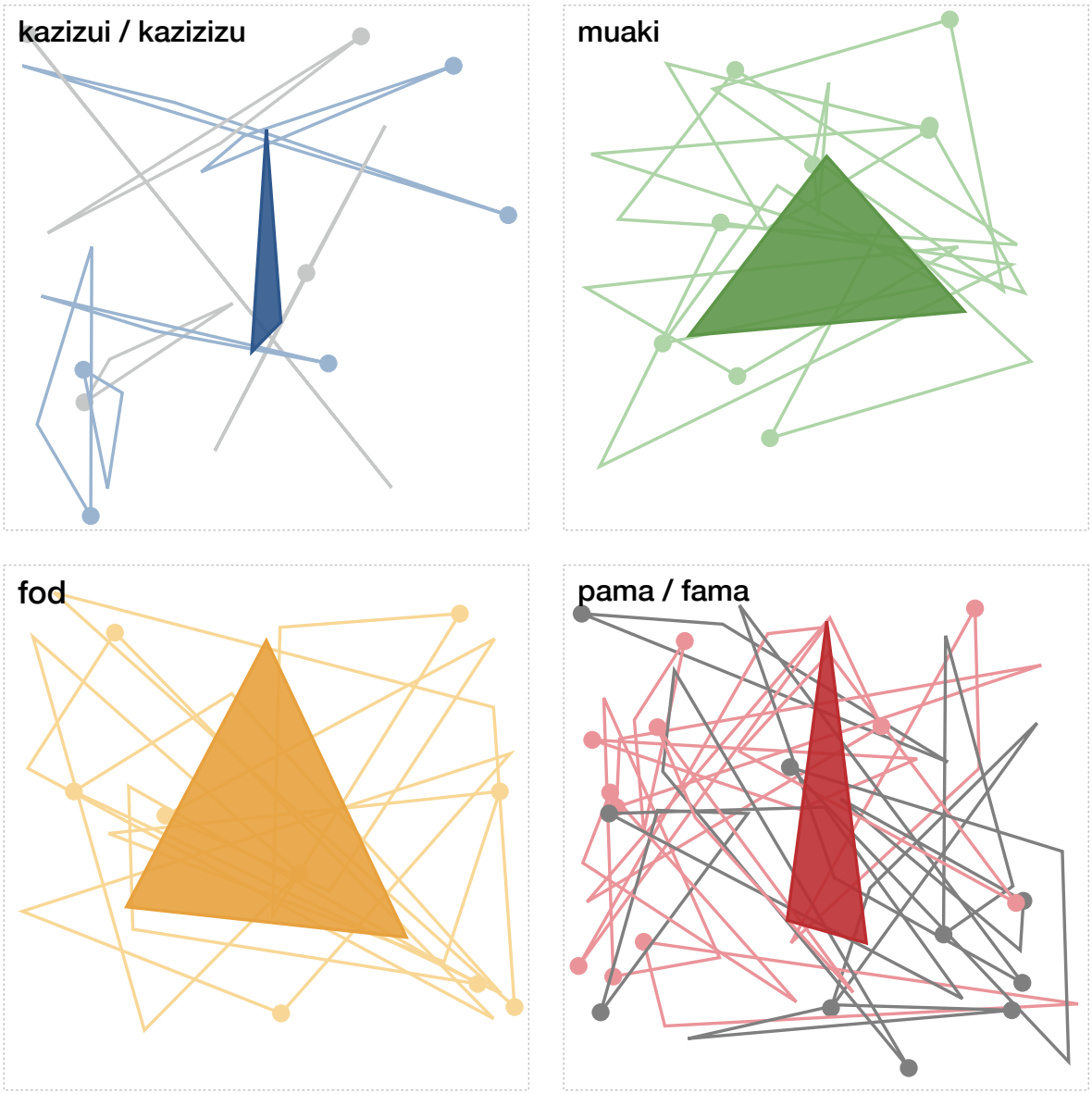


Figure 3.6 The four main categories that emerged in chain A by generation 9. The triangles in the background are those that were assigned a given name, while the filled triangles in the foreground are prototypes of those triangles which were produced by averaging the triangles together. Colours correspond to those in Fig. 3.4.

of words, with *muaki* denoting the more equilateral triangles and *fama* / *pama* denoting those with less even sides. The distinction between *fama* and *pama* is unclear and may in fact be non-existent; the participant noted: '*fama* and *pama* seemed to be related in some way … but couldn't quite figure out how'. The previous participant, A8, claimed that *pama* and *fama* were used for 'non-equilateral triangles', but the distinction between the words remains unclear for most of their lineage. If we consider, *pama* and *fama* to be variants on the same category, then we have four categories in this language.

As in chains A and B, the language in chain C (at generation 8) divides the triangles up into small, medium, and large. Small and thin triangles are referred to as *kik*, medium-sized

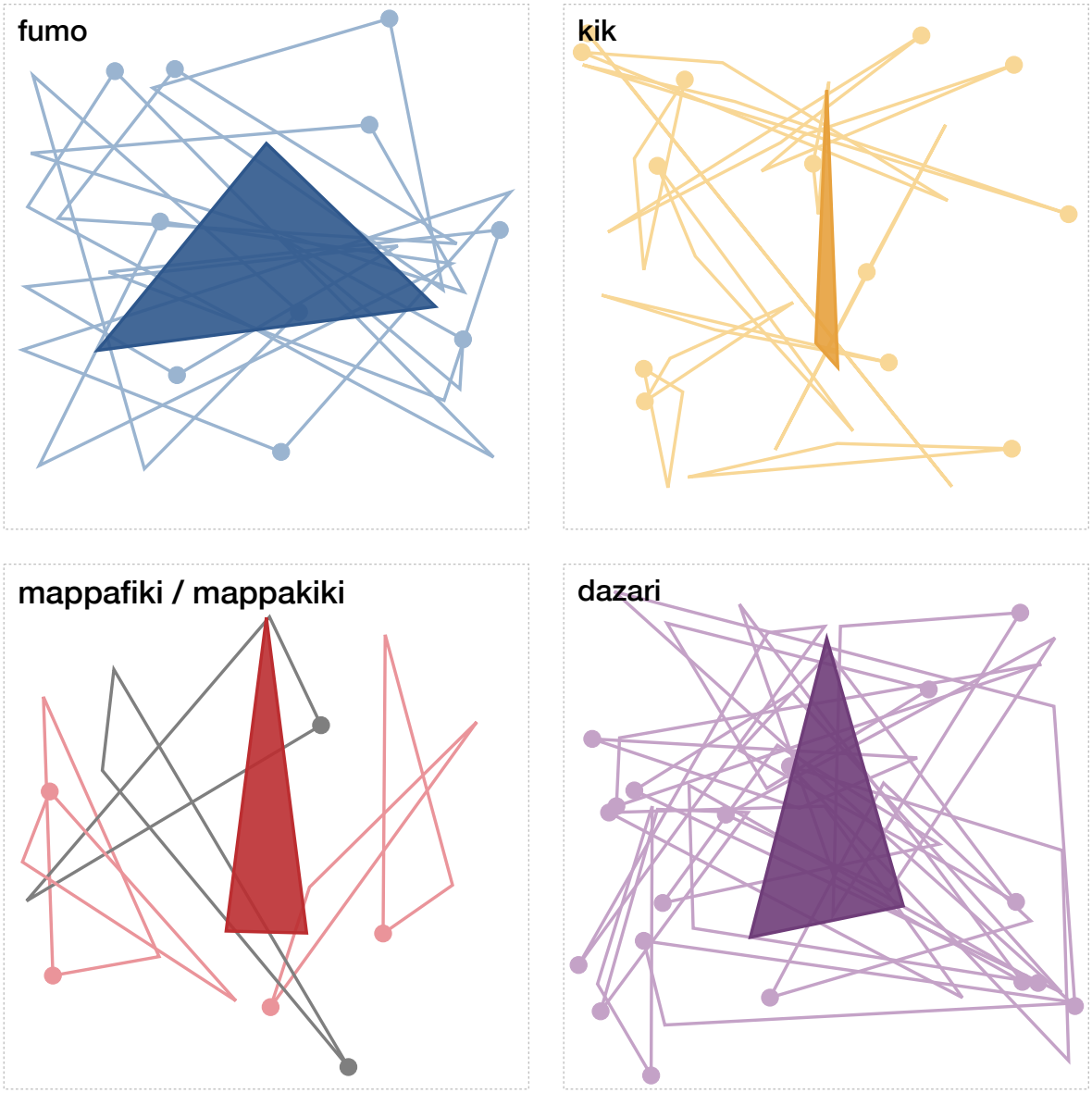


Figure 3.7 The four main categories that emerged in chain C by generation 8. The triangles in the background are those that were assigned a given name, while the filled triangles in the foreground are prototypes of those triangles which were produced by averaging the triangles together. Colours correspond to those in Fig. 3.4.

triangles are called *dazari*, and large triangles are *fumo*. There are two exceptions to the *dazari* category: *mappakiki* and *mappafiki* both of which appear to cover medium-sized but relatively pointy triangles[25]. Again, the distinction between the *mappa-* forms is unclear and if we consider them part of the same category, then there are also a total of four categories in this language. As mentioned previously, Chain D resisted the emergence of categorical structure, as can be seen from the plot in Figure 3.4; as early as generation 5, a single word was being used for the vast majority of triangles.

[25] Note that *dazarai* and *mappfiki* are most likely typos on *dazari* and *mappafiki*.

3.2. Experiment 2

Forty participants (25 female) completed experiment 2 during the period 8th—19th July 2013. The mean age was 23.4 years ($SD = 5.02$). There was no significant correlation between generation number and the time of the day at which the experiment was completed ($r = -0.038$, $n = 40$, $p = 0.815$). Generation number was also not significantly correlated with age ($r = 0.25$, $n = 39$, $p = 0.124$; one anomaly removed), sex ($r = 0.027$, $n = 40$, $p = 0.869$), or language experience ($r = -0.191$, $n = 40$, $p = 0.238$). However, it was significantly correlated with the day of the week on which the experiment was completed ($r = 0.444$, $n = 40$, $p = 0.004$).

3.2.1. Number of unique strings

Unlike experiment 1, the number of unique strings is not able to collapse. In fact, the number of words in the DYNAMIC SET never falls below 22 (16 is the minimum for $\lambda = 3$), and by generation 10, there are an average of 28 words in each of the languages. In addition, the number of unique strings in the STABLE SET also remains high, despite there being no limit on the number of times a word could be repeated within this set. These results are illustrated in Figure 3.8. However, although the number of unique strings does not collapse as dramatically as in experiment 1, there is still a significant downward trend, in both the DYNAMIC SET ($L = 1912.5$, $m = 4$, $n = 11$, $p < 0.001$) and the STABLE SET ($L = 1908$, $m = 4$, $n = 11$, $p < 0.001$).

3.2.2. Learnability

The results for transmission error are illustrated in Figure 3.9.A. The graph suggests

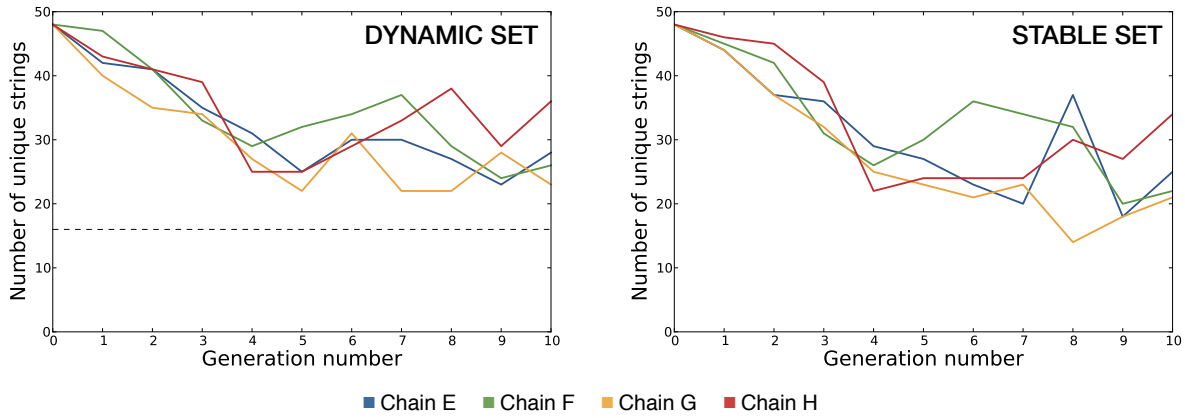


Figure 3.8 Number of unique strings over the 10 generations in chains E–H for the DYNAMIC SET and STABLE SET. Unlike experiment 1, the languages do not collapse to a small set of words. The dashed line shows the minimum number of unique strings (16) that could be used in the dynamic set due to the λ constraint.

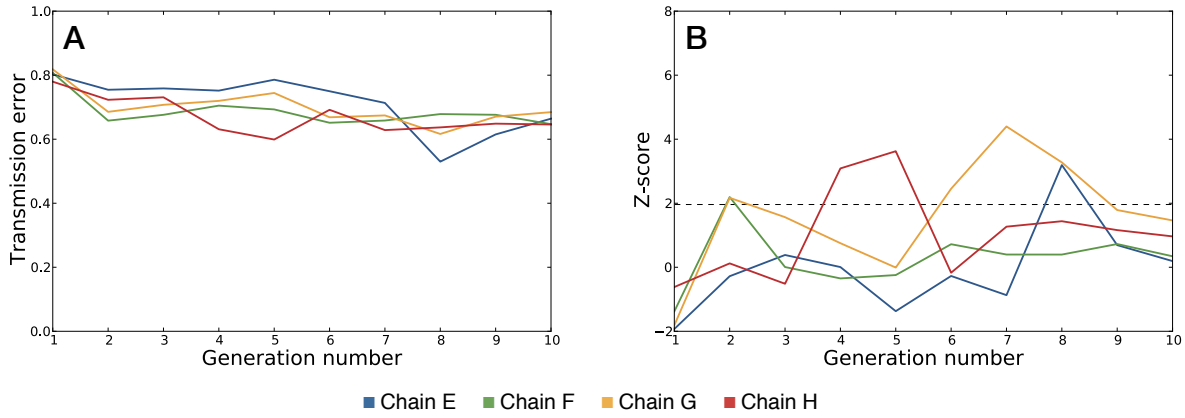


Figure 3.9 (A) Transmission error over the course of 10 generations for chains E—H. **(B)** *Z*-score transformation of the error scores by comparing the veridical score to a distribution of error scores for Monte-Carlo permutations of the stable word set. The dashed line in (B) show the two-tailed 95% confidence level.

that error in experiment 2 remained relatively static over the 10 generations. Nevertheless, the results do show a significantly decreasing trend ($L = 1415$, $m = 4$, $n = 10$, $p < 0.001$). To account for the possibility that transmission error may be decreasing simply because chance guesses are increasingly more likely to be correct, the transmission error scores were subject to the same transformation as described in Section 3.1.2 and are shown in Figure 3.9.B. These results suggest that, in many cases, participants are performing no better than chance, although there are a small number of participants whose error scores are significantly non-random. Nevertheless, there is a significant upward trend ($L = 1351$, $m = 4$, $n = 10$, $p = 0.005$).

The amount of error in the mini tests decreased significantly over the course of the experiment ($L = 1312$, $m = 4$, $n = 10$, $p = 0.032$), falling from an average of 35.9% error in generation 1 to an average of 19.1% error in generation 10. There was no significant decrease in the time spent on each test item ($L = 1202$, $m = 4$, $n = 10$, $p =$ n.s.) nor participants' difficulty ratings ($L = 1266$, $m = 4$, $n = 10$, $p = 0.154$).

3.2.3. Structure

The structure results for all eight triangle distance metrics are given in Figure 3.10. As in experiment 1, the Bonferroni correction has been applied to adjust for the fact that eight separate metrics have been used to quantify the structure in the languages, raising the confidence level to $z = 2.734$. Table 3.2 gives the mean and standard deviation of the z-scores for each metric.

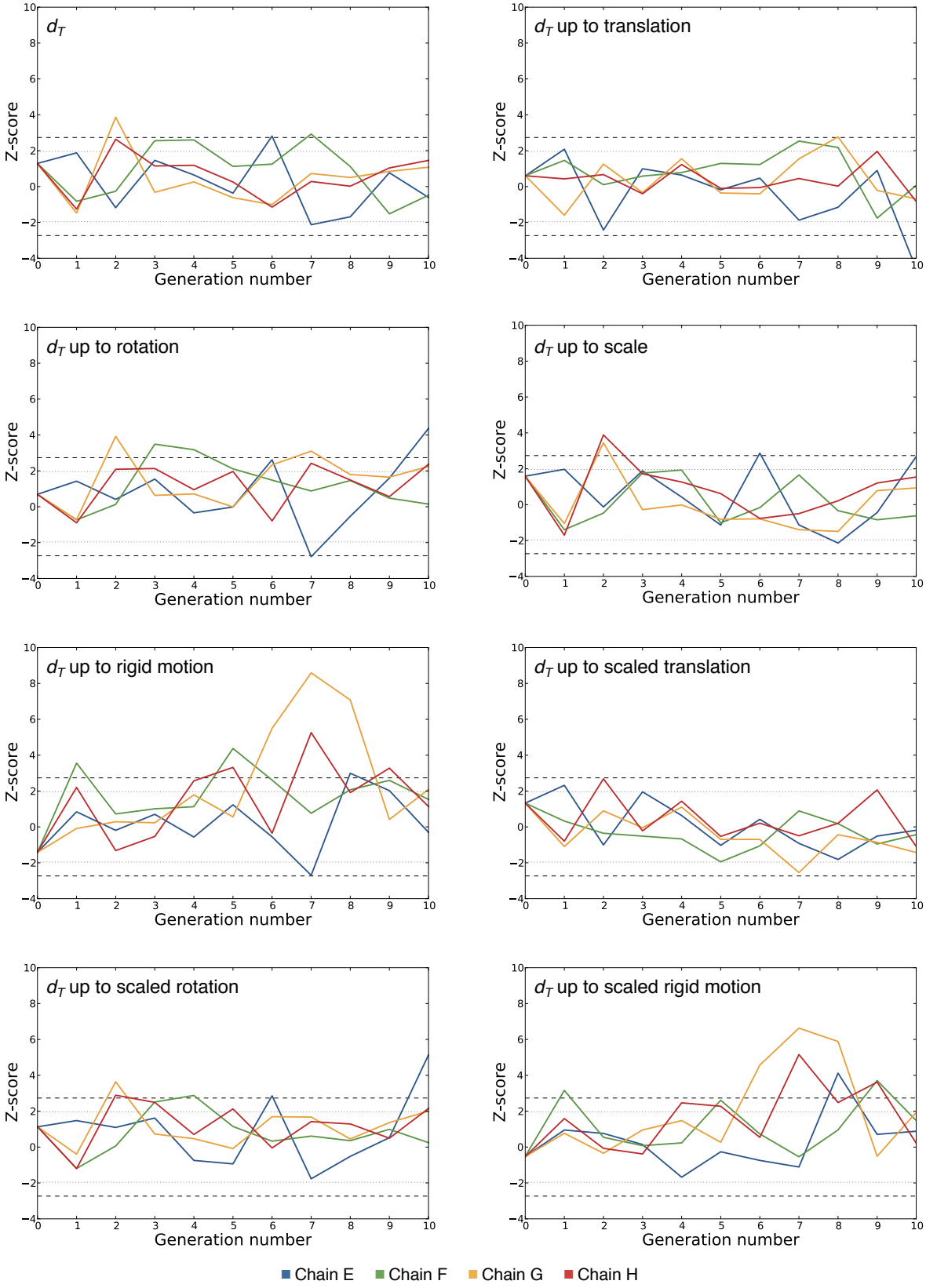


Figure 3.10 Plots showing the structure in the languages under the eight triangle distance metrics. The dotted lines show the *z*-scores equivalent to $\alpha = 0.05$ and the dashed lines show the *z*-scores equivalent to $\alpha = 0.00625$ which corrects for eight multiple comparisons using the Bonferroni correction. Scores that lie outside of the dashed lines are significantly unlikely to have occurred by chance alone.

Table 3.2 Mean and standard deviation of the structure *z*-scores for each distance metric

Distance metric	Mean	SD
d_T	0.561	1.374
d_T up to translation	0.286	1.361
d_T up to rotation	1.171	1.407
d_T up to scale	0.419	1.465
d_T up to rigid motion	1.41	2.284
d_T up to scaled translation	–0.034	1.174
d_T up to scaled rotation	1.025	1.34
d_T up to scaled rigid motion	1.231	1.915

Again, the structure scores that stand out are those for $d_{T_{rm}}$ and $d_{T_{srm}}$. Furthermore, there is a significant upward trend for both $d_{T_{rm}}$-based scores ($L = 1780$, $m = 4$, $n = 11$, $p = 0.002$) and $d_{T_{srm}}$-based scores ($L = 1758$, $m = 4$, $n = 11$, $p = 0.006$)[26]. These results reiterate the idea that the experiments are being driven by a strong bias for shape and size. To investigate these properties more closely, we produced shape-size scatterplots for the most structured generation in each chain (according to the $d_{T_{rm}}$ results). However, since each word in the STABLE SET is used just 1.8 times on average, it becomes difficult to visualize any emergent categorical structure. To help us visualize the structure, we used agglomerative hierarchical clustering[27] to cluster the strings in the STABLE SET. Determining the number of clusters c in a dataset is a classic computational problem (Duda, Heart, & Stork, 2001, p. 557). Rather than select some algorithm to determine the number of clusters, we set c to 5, since the shape-size space in experiment 1 was typically divided into around four or five categories. The plots are shown in Figure 3.11.

The way in which the spaces are carved up is similar to how they were divided in experiment 1. Typically, there appear to be three main categories – small, medium, and large – although the boundaries appear to be less well defined. Since G7 had a structure score on par with those in experiment 1, we took a look at this language in particular. The triangles represented by the five clusterings are shown in Figure 3.12. *Kik* covers the small triangles and cluster 2 (*do-* forms) covers the large triangles. Medium-sized triangles are referred to by cluster 3 (*-no* forms) and cluster 4 (*-zu* forms). Cluster 5 (*-iki* forms) tends to be very pointy.

---

[26] After excluding generation-0 data points, d_{Trm} remains significant ($L = 1303$, $m = 4$, $n = 10$, $p = 0.045$), while d_{Tsrm} becomes non-significant ($L = 1298$, $m = 4$, $n = 10$, $p = 0.055$).

[27] Distance metric: normalized Levenshtein distance; linkage criterion: mean.

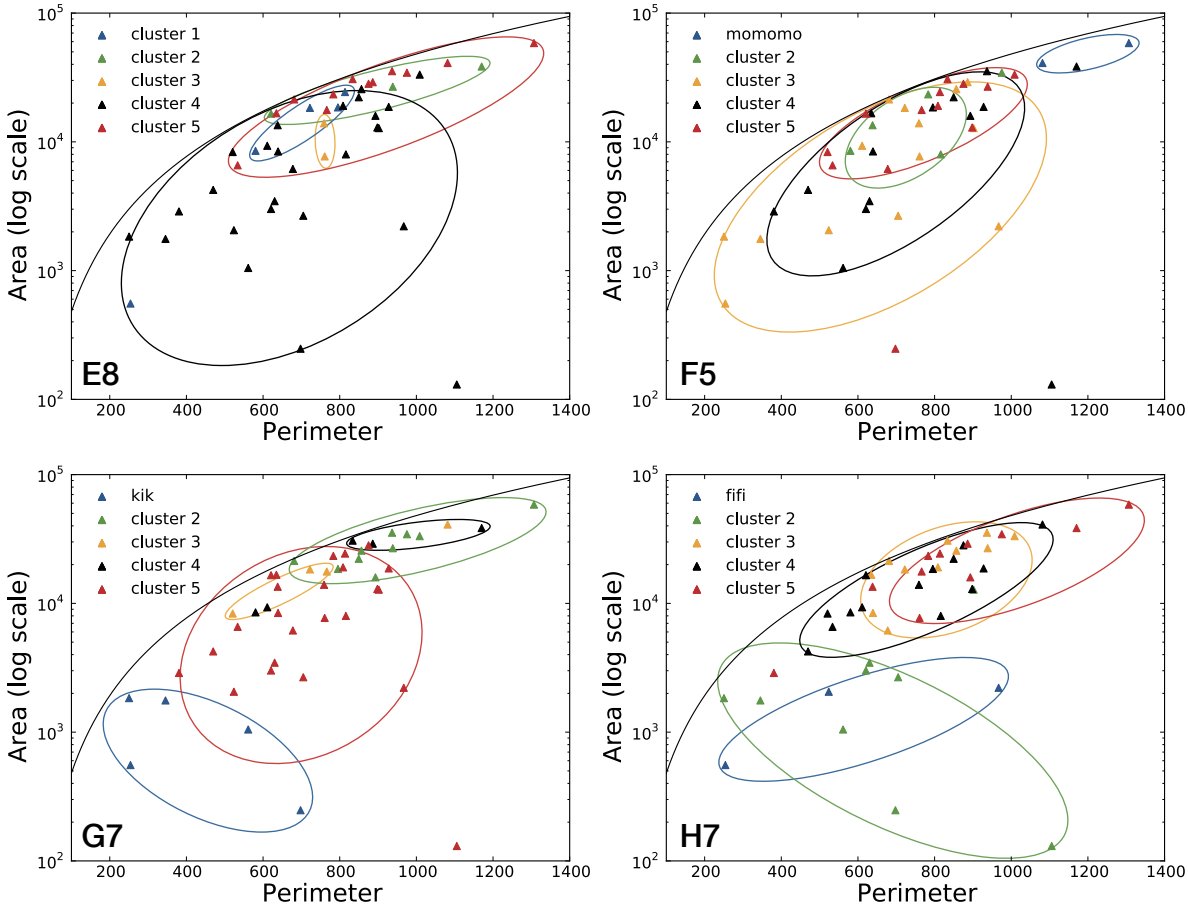


Figure 3.11 Scatterplots showing perimeter against area for the 48 triangles in the STABLE SET for four participants. Points represent triangles and are colour-coded according to the cluster of similar strings that the word used to describe that triangle belongs to. The space inside the plots roughly corresponds to shape vs. size. Black curves show the relationship between area and perimeter for perfectly equilateral triangles. As you move along the curve, the triangles go from small to large; as you move away from the curve, the triangles become skinnier. Ellipses show interpretations of how the space may be carved up. **E8** cluster 1 = {zuma, zupa}, cluster 2 = {duh, dupa, dus}, cluster 3 = {fuk, fuki}, cluster 4 = {datiki, fahliki, falihki, fatiki, fokiki, folihki, fotiki, kiki, liki, mahiki, pofiki, pokiki, polihki, poliki, pontiki, potiki, taliki, tiki, tokiki, zokiki, zukiki}, cluster 5 = {fokaki, folaki, mofiki, mohaki, mohalifi, momahiki, momaki, momiki, pohaki}. **F5** cluster 2 = {modazi, mudazi}, cluster 3 = {duzaki, mukaka, mukaki, mukazu, muzaka, zakaka, zkaka, zukaka}, cluster 4 = {kimiku, kizu, maziki, mizaku, miziki, miziku, mizu, mizumi, mozimu, zimu}, cluster 5 = {duma, kima, zakami, zikiki, zikimo, zikumo, zimomo, zukimi, zuma}. **G7** cluster 2 = {dod, domo, domod}, cluster 3 = {mino, reno}, cluster 4 = {kikioluazu, kikoluazu, moazu, padzu, panzu}, cluster 5 = {demiki, demikiki, kiki, mafiki, maziki, miki, mikiki, rafiki, raifiki, raifkiki, zafiki, zaifiki}. **H7** cluster 2 = {akiki, kiki}, cluster 3 = {akuzo, azizo, azuzo, mazuzo, muzuzo, zizu, zuzi}, cluster 4 = {atzuki, azuki, matatiki, matziki, matzuki, matzuma, matzuzi, maziki, mazuzi}, cluster 5 = {akuma, azima, fima, puma, putafima}.

The purpose of experiment 2, however, was to allow for the emergence of compositionality, but by clustering the words into small groups, we necessarily lose any compositional structure that might have been present. With this in mind, we took a closer look at one particular cluster of words in G7. Within cluster 5 (*-iki* forms), there are four distinct groups beginning *de-*, *ma-* or *mi-*, *ra-*, and *za-*. Figure 3.13 shows the triangles within each of these four subgroups. When we separate out the triangles in this way, new structure emerges. The *de-* and *ra-* groups are clearly wider, while the *m-* and *za-* groups are

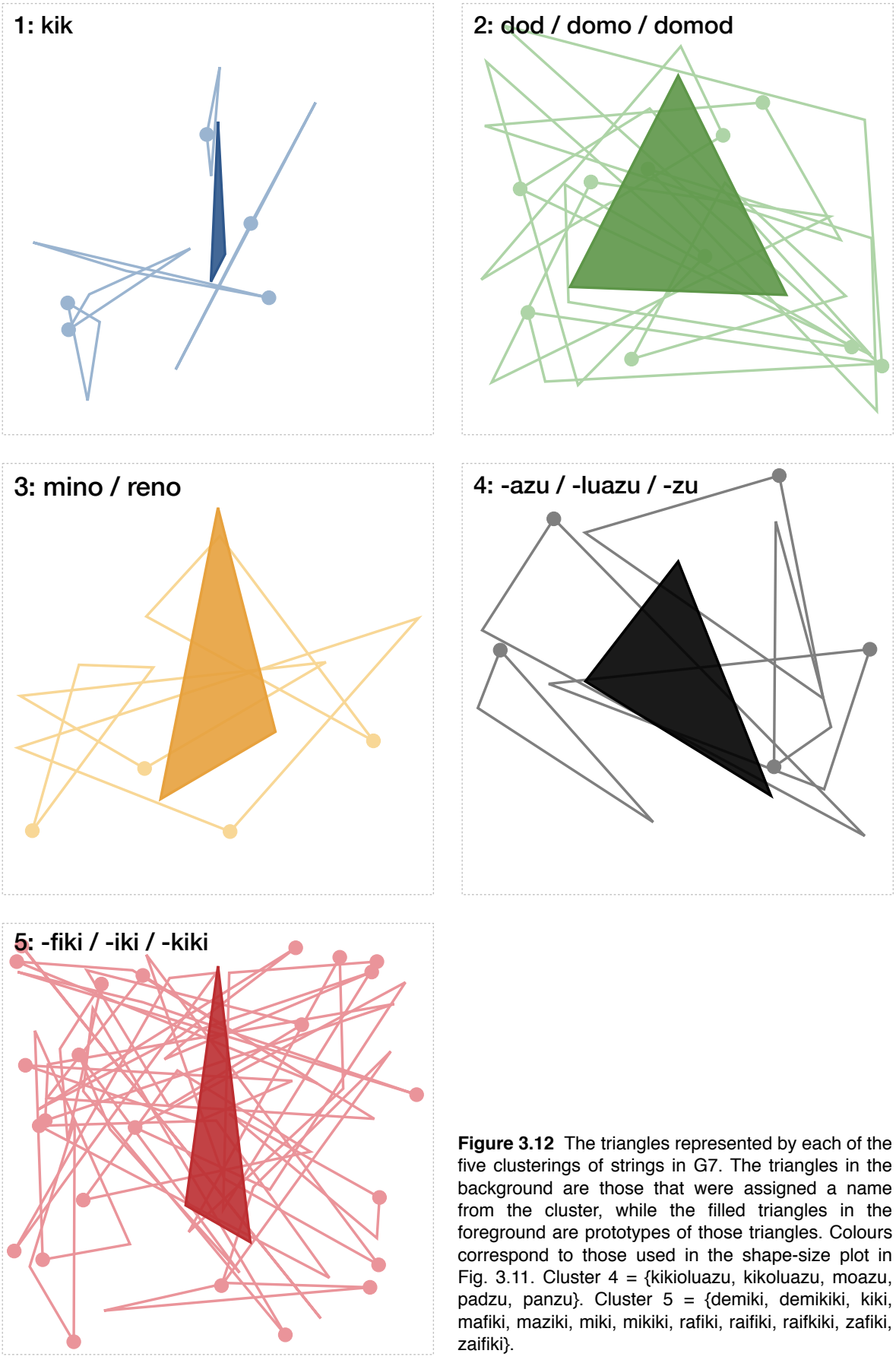


Figure 3.12 The triangles represented by each of the five clusterings of strings in G7. The triangles in the background are those that were assigned a name from the cluster, while the filled triangles in the foreground are prototypes of those triangles. Colours correspond to those used in the shape-size plot in Fig. 3.11. Cluster 4 = {kikioluazu, kikoluazu, moazu, padzu, panzu}. Cluster 5 = {demiki, demikiki, kiki, mafiki, maziki, miki, mikiki, rafiki, raifiki, raifkiki, zafiki, zaifiki}.

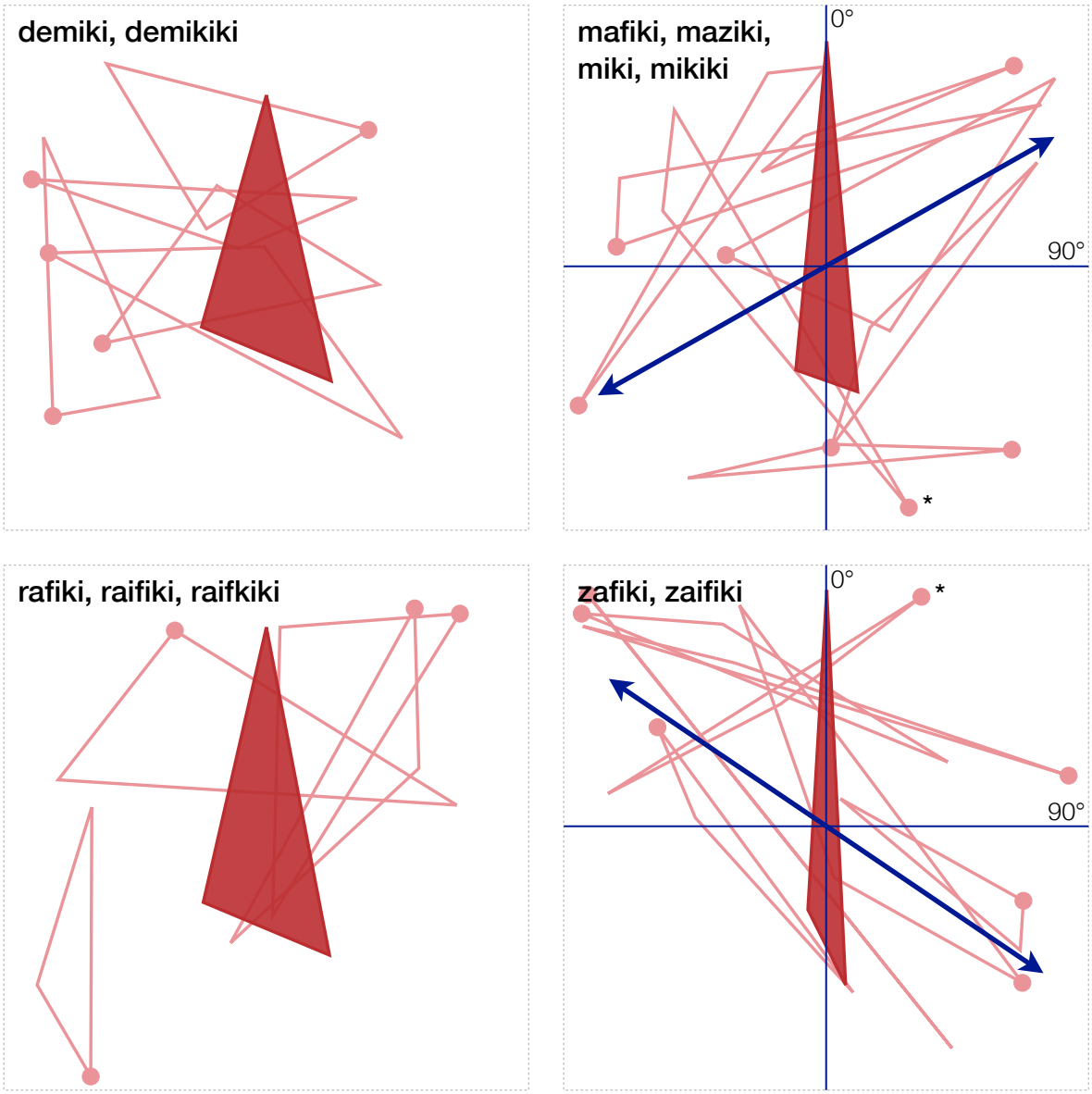


Figure 3.13 Cluster 5 in G7 subdivided into four groups according to first syllable. Within this cluster, we see signs of substructure. *de-* and *ra-* words correspond to wider triangles, and *m-* and *za-* words correspond to thiner triangles. The *m-* words tend to point NE–SW, while the *za-* words tend to point NW–SE. Blue arrows show the mean angles of rotation (excluding the anomaly in each set marked with an *).

thinner. Interestingly, there appears to be some rotational structure. The *m-* group has a mean angle of rotation of 72.6°, while the *za-* group has a mean angle of rotation of 114.4°. To confirm that this was not simply chance, we looked at the same groups of words in the participant's DYNAMIC SET which had mean angles of rotation of 52.8° and 121.6° respectively. Taking the data from the STABLE and DYNAMIC SETS combined, the mean angles of rotation were 61.4° (95% CI: 44.3°—78.7°) for the *m-* group and 117.3° (95% CI: 103.3°—131.5°) for the *za-* group. The difference between the groups is significant ($t = 4.524$, $df = 26$, $p < 0.001$), suggesting that they do indeed mark out a distinction in

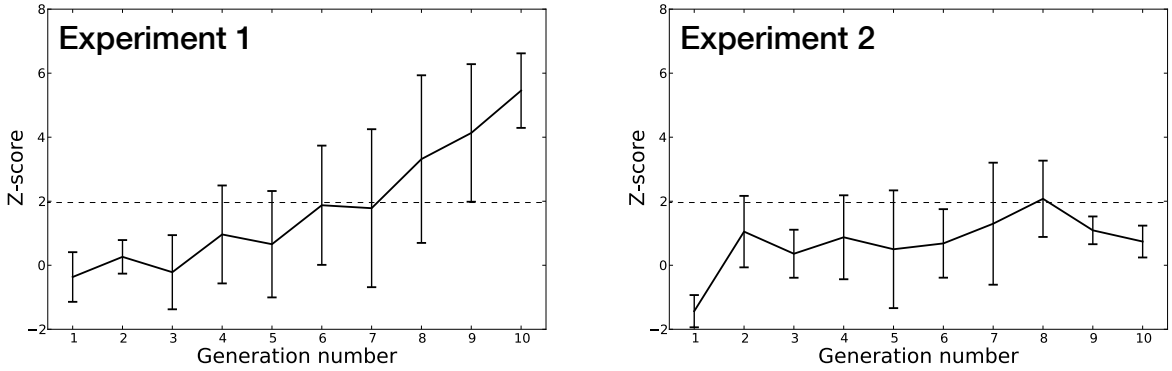


Figure 3.14 Mean learnability scores for experiment 1 and experiment 2. Error bars show the 95% confidence intervals on the mean and the dashed lines show the 95% confidence level.

rotation[28]. However, we cannot conclude from this that the *ma-/mi-* and *za-* syllables are morphemes, since there is no evidence of them being used productively elsewhere in the language. The syllable *mi-* does occur in the word *mino* in cluster 3, but there was no evidence that *mino*s were angled the way you would expect if *mi-* were productive.

3.3. Comparison of experiments 1 and 2

In this section, we briefly compare the outcomes of the two experiments.

3.3.1. Learnability

Figure 3.14 plots the mean learnability scores for both experiments. In the first five generations, the plots reveal a similar trend. It is only after generation 5 that we see the two experiments diverge on separate trajectories. This point of divergence corresponds closely with the collapse in the number of unique strings we observed at around generation 5 in experiment 1, suggesting some interaction here.

3.3.2. Structure

Figure 3.15 plots the mean $d_{T_{rm}}$-based structure scores for both experiments[29]. As with the learnability results, the two experiments follow similar trajectories up until around generation 5, at which point the two experiments diverge with experiment 1 giving rise to more structured languages that are highly non-random. These results were supported by a qualitative analysis of the languages, which showed that systematic category structure

[28] Independent samples *t*-test; includes the two anomalies pointed out in Fig. 3.13. Also significant if you consider the STABLE SET ($t = 2.236$, $df = 12$, $p = 0.045$) or DYNAMIC SET ($t = 4.097$, $df = 12$, $p = 0.001$) alone.

[29] As with the analyses above, we focus only on d_{Trm} here, since this metric appears to best capture the structure in the languages.

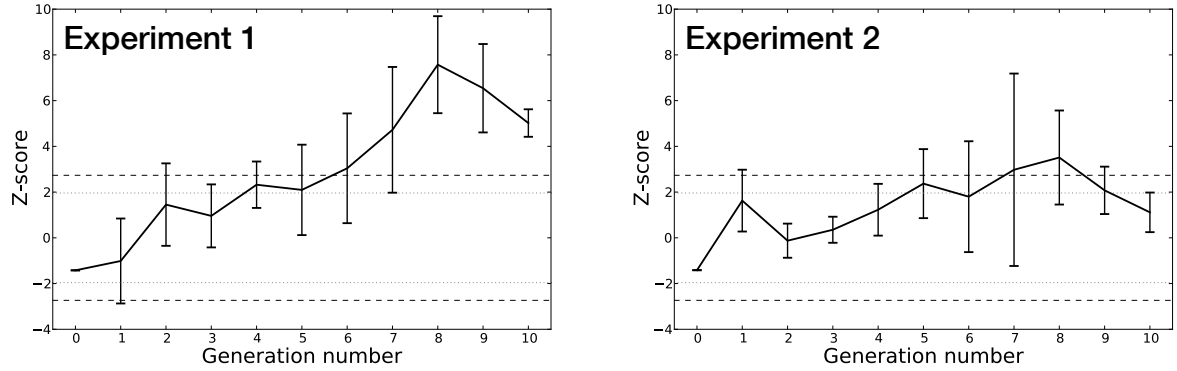


Figure 3.15 Mean $d_{T_{rm}}$-based structure scores for experiment 1 and experiment 2. Error bars show the 95% confidence intervals on the mean and the dashed lines show the 95% confidence level for significantly structured languages (Bonferroni corrected for 8 multiple comparisons).

emerged in experiment 1 and, to a certain extent, in experiment 2. However, there was no evidence that the procedural modification in experiment 2 promoted the evolution of compositional linguistic structure.

3.3.3. Intergenerational correlation between structure and learnability

Hypothesis III was that we could show a link between structure and learnability. The correlation between the learnability scores of generation i and the $d_{T_{rm}}$-based structure scores for the DYNAMIC SET[30] of generation $i-1$ was significant for experiment 1 ($r = 0.786$, $n = 36$, $p < 0.001$) and experiment 2 ($r = 0.37$, $n = 40$, $p = 0.019$). Nevertheless, it is unsurprising that these variables are correlated given that we have already shown empirically that structure and learnability increase over the course of the experiments. We therefore ran a partial correlation between these variables controlling for generation number[31]. This test was significant for experiment 1 ($r = 0.479$, $n = 36$, $p = 0.002$) and experiment 2 ($r = 0.307$, $n = 40$, $p = 0.0498$), suggesting that structure is, at least in part, driving the increase in learnability.

The code used to produce the results presented in this chapter can be found in `analysis.py` (Appendix F). In the following chapter, we discuss how the results from these experiments relate to the original research questions and hypotheses.

---

[30] We use the DYNAMIC SET here, since we want to correlate a participant's learnability score with the structure score of the training material that participant received, not the general structure score of the previous participant.

[31] This approach was adopted from Seyfarth (2010, p. 22). It can be thought of in the following way: if generation number were to have no effect on structure and learnability, would there still be a correlation between participants' learnability scores and the structure scores of their training materials? This is one simple way of controlling for the non-independence of the scores.

4. Discussion

In this final chapter, we discuss how the empirical results relate to the research questions. Section 4.1 summarizes the general findings. Section 4.2 considers alternative mechanisms to explain the increase in learnability. Section 4.3 discusses the contributions this project has made. Section 4.4 offers some criticisms. Section 4.5 concludes with some thoughts on future research directions.

4.1. General findings

Table 4.1 summarizes the results in terms of the hypotheses. Hypothesis I was supported in both experiments: the languages evolved to become more learnable. Hypothesis II was also supported: the emergence of categorical structure allows participants to accurately predict how the triangles in the STABLE SET had previously been labelled by exploiting the regularity in the mapping between linguistic form and arbitrarily-defined regions of the meaning space. Finally, Hypothesis III was supported suggesting that the emergence of structure is driving increases in learnability.

Table 4.1 Summary of the outcomes of experiments 1 and 2

Hypothesis	Test statistic	Experiment 1	Experiment 2
I: Learnability increases with generation number	Page's trend test	$p < 0.001$	$p = 0.005$
II: Structure increases with generation number	Page's trend test	$p < 0.001$	$p = 0.002$
III: Increased structure explains increased learnability	Partial correlation	$p = 0.002$	$p = 0.0498$

It is worth thinking about the mechanism underlying these experiments. Firstly, the number of words in the languages decreases due to the bottleneck on transmission. Participants cannot memorize as many as 48 words within their 15- to 20-minute training phase, let alone the mappings between words and meanings. Consequently, at every iteration the number of words drops, eventually stabilizing on a small, memorable set. Secondly, as the number of words in the languages decreases, correspondences between form and meaning arise by chance. Since participants expect to find regularity in the

languages[32], they infer rules to explain such chance correspondences. This leads to emergent global structure as each subsequent participant imposes additional local structure. Thirdly, as structure increases, the languages become more learnable because participants can generalize from the rules they have inferred to correctly identify items they never saw during training.

In experiment 2, we impede this three-step mechanism by preventing the variation in the languages from collapsing. We predicted that the languages would find another way to circumvent the bottleneck, perhaps by evolving compositionality as previously observed by Kirby et al. (2008). However, this was not supported by our results. There are at least two possible explanations for this. Firstly, it is possible that the languages in experiment 2 were not given sufficient time to evolve a compositional system. This could be tested by extending the chains. Secondly, it is possible that the artificial constraint does not really enforce expressivity in the same way that the functional use of language enforces expressivity. This could be tested by implementing a dyadic version of the experiment in which participants negotiate an optimal system in a communication task.

This is something that we considered during initial discussions but ruled out for two reasons. Firstly, since this was the first time an infinite meaning space had been implemented in an iterated learning experiment, we preferred to keep the experiment as simple as possible. Secondly, the use of dyads scuppers what is arguably the most important aspect of iterated learning: the fact that no participant is consciously trying to design an optimal system. Each participant in a monadic diffusion chain is tasked with learning and recalling the language presented to them. We even tell them that a prize will be awarded for the participant best able to recall the language. Despite this, the outcome we have consistently observed is systematic structure that is emergent from the process of iterated learning itself and therefore akin to Keller's (1994) 'invisible hand'.

4.2. Alternative mechanisms for increased learnability

The emergence of categorical or compositional structure is not the only way in which transmission fidelity could be improved. Here we investigate two alternative sources of increased learnability: combinatorial structure and sound symbolism.

[32] For example, on completing the experiment, one participant (F5) claimed that the language she was learning was wrong and contained errors and inconsistencies.

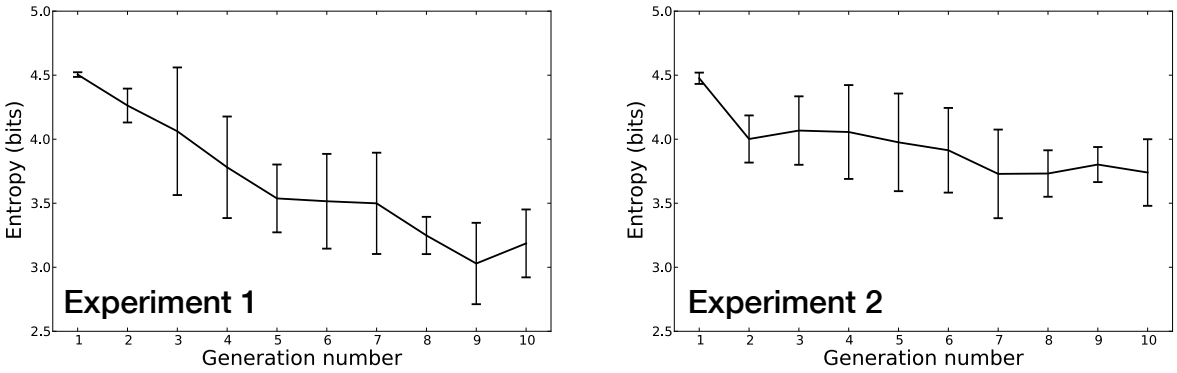


Figure 4.1 Mean entropy of participants' training materials (i.e. the DYNAMIC SET of the previous participant) for experiment 1 and experiment 2. Error bars show the 95% confidence interval on the mean.

4.2.1. Combinatorial structure

One possible explanation is that predictable combinatorial structure evolved during transmission. The information theoretic measure of Shannon entropy (Shannon, 1948) can be used to estimate the predictability of a language. To calculate this, we determine the probability of occurrence p for each syllable s in the set of syllables in a participant's training material S, and compute the entropy H according to

$$H(S) = -\sum_{s \in S} p(s) \cdot \log_2 p(s). \tag{4.1}$$

The results are given in Figure 4.1. There is a significant downward trend in experiment 1 ($L = 1140$, $m = 3$, $n = 10$, $p < 0.001$) and experiment 2 ($L = 1408$, $m = 4$, $n = 10$, $p < 0.001$). The partial correlation between participants' learnability scores and the entropy of their training materials (controlling for generation number) was significant for experiment 1 ($r = -0.388$, $n = 30$, $p = 0.029$), but not for experiment 2 ($r = -0.053$, $n = 40$, $p = 0.746$). This suggests that decreasing entropy may explain some of the increase in learnability in experiment 1.

Calculating the entropy of a syllable set informs us about the predictability of the system as a whole, but not whether one unit is predictable in light of another. Therefore, we calculated the conditional entropy of participants' training materials, which measures the average predictability of the second syllable $y \in Y$ given that the first syllable $x \in X$, for all bi-syllables, is known. It is given by,

$$H(Y|X) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \cdot \log_2 p(y|x). \tag{4.2}$$

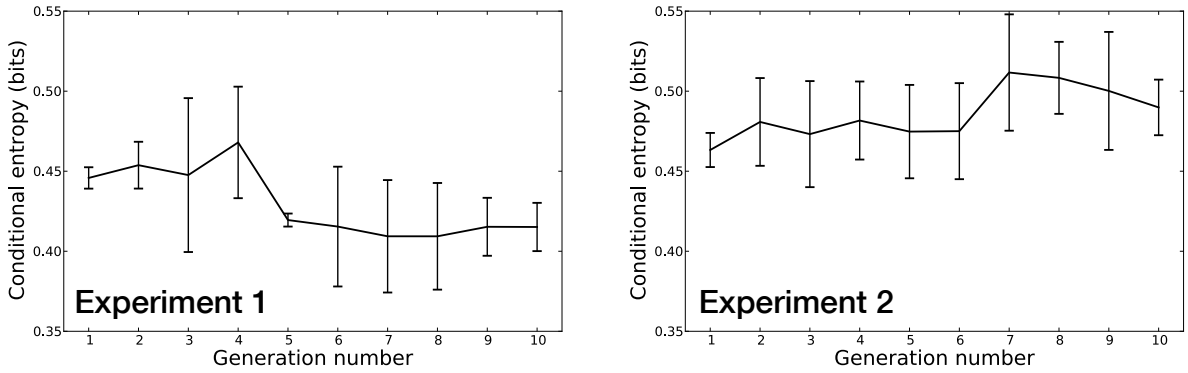


Figure 4.2 Mean conditional entropy of participants' training materials (i.e. the DYNAMIC SET of the previous participant) for experiment 1 and experiment 2. Error bars show the 95% confidence interval on the mean.

The results are given in Figure 4.2. There was a significant downward trend in experiment 1 ($L = 1012$, $m = 3$, $n = 10$, $p = 0.014$), but not in experiment 2 ($L = 1116$, $m = 4$, $n = 10$, $p =$ n.s.). The partial correlation between participants' learnability scores and the conditional entropy of their training materials (controlling for generation number) was not significant for experiment 1 ($r = 0.277$, $n = 30$, $p = 0.134$) or experiment 2 ($r = 0.212$, $n = 40$, $p = 0.187$).

4.2.2. Sound symbolism

Sound symbolism refers to the phenomenon where a unit of sound 'goes beyond its linguistic function as a contrastive, non-meaning-bearing unit, to directly express some kind of meaning' (Nuckolls, 1999, p. 228). The qualitative analysis suggested that sound symbolism might play a role in the languages, which is particularly relevant here as it has been suggested that sound symbolism facilitates word learning (see e.g. Monaghan, Christiansen, & Fitneva, 2011; Nygaard, Cook, & Namy, 2009; Parault & Schwanenflugel, 2006). The emergence of sound symbolic patterning was explored in the following way: for each of the ten phonemes that initialized the experiments, we measured the mean pointedness of triangles whose associated words contained that phoneme. The pointedness of a triangle $\mathcal{T}$ was calculated according to $\frac{\log_e(\text{perimeter}(\mathcal{T})^2/(12\sqrt{3}))}{\log_e(\text{area}(\mathcal{T}))}$, i.e. the ratio of the log of the triangle's area to the log of the area of an equilateral triangle with the same perimeter. The results are given in Figure 4.3. In experiment 1, the phonemes [iː], [k], and [z] tend to be associated with more pointed triangles, while the phonemes [əʊ], [d], [f], and [m] tend

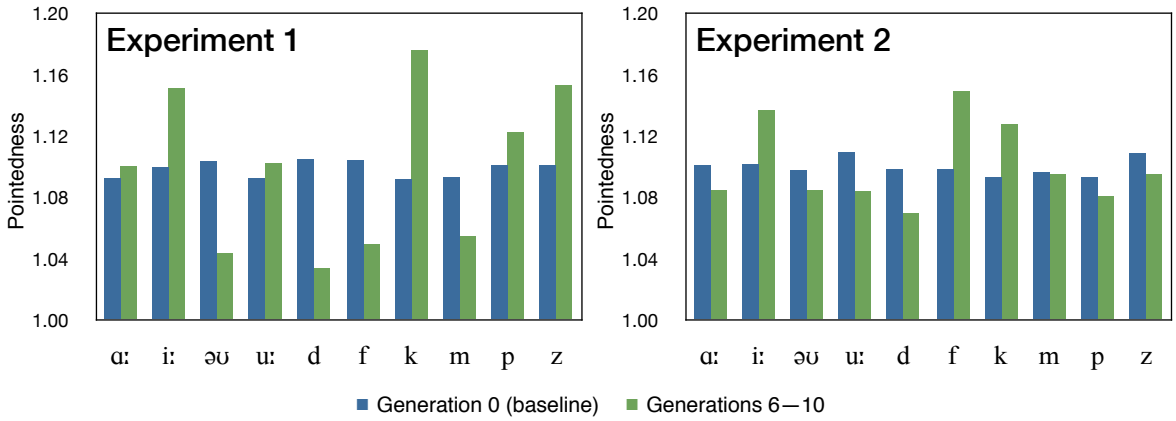


Figure 4.3 Mean pointedness of triangles that correspond to words containing at least one of the ten phonemes that initialized the experiments. The blue bars give the score for generation 0 data across all chains, which provides a baseline, since these words were generated randomly and there should be no sound symbolic correspondence. The green bars give the score for generations 6—10 across all chains (excluding D).

to be associated with more equilateral triangles[33]. Similar findings hold for experiment 2, although the effect is weaker. We have not conducted statistical analyses on these patterns, since the nature of the data does not lend itself to such analysis. However, the emergence of attested sound symbolic patterns may have made some contribution to increased learnability.

4.3. Contributions

These experiments have shown that categorical structure can emerge in an iterated learning experiment with a meaning space that is infinite in magnitude. We describe our meaning space as infinite because it avoids the four characteristics of fixed, monolithic spaces. Firstly, the meaning space is **not pre-specified**. We had no particular hypothesis about which features participants would find salient. For this reason, it was necessary to construct eight separate distance metrics to cover the most likely possibilities. Secondly, the meaning space is **not discrete**. On each dimension, the triangle stimuli vary over a continuous scale. In the size dimension, the triangles vary from 1 pixel to thousands of pixels; in the rotation dimension, the triangles vary from 0° to $359.\overline{9}$°; in the shape dimension, the triangles may be equilateral, isosceles, scalene, right-angled, or anywhere in between; and in the location dimension, the triangles may be located anywhere in the plane. Thirdly, the meaning space is **not finite**. The number of stimuli is limited by the

[33] These relationships coincide with the literature, which claims that voiceless stops (e.g. [k]) and closed unrounded vowels (e.g. [iː]) tend to be perceived as angular, while voiced stops (e.g. [d]), sonorants (e.g. [m]), and open rounded vowels (e.g. [əʊ]) tend to be perceived as more rounded (Ahlner & Zlatev, 2010, p. 310).

density of the pixels on the display and ultimately by the resolution of the human retina. However, even with this limit, the number of possible triangles is 6×10^{15}, which is infinite as far as the human mind is concerned. Fourthly, the meaning space is **not low-dimensional**. The obvious dimensions are shape, size, orientation, and location, but the systems are not constrained by these four dimensions alone. In fact, there are certain dimensions that none of our metrics consider, such as reflection, rotation based on some property other than the orienting spot, or shape-invariant transformations. While this makes the experiments methodologically difficult, it demonstrates that they meet one trait of a truly infinite space: *a-priori* unpredictability of the quality dimensions. The implementation of an infinite meaning space is the primary contribution that this dissertation has made.

In order to analyse the data, we needed to devise various novel techniques. These included: a measure of distance between two triangles given that one vertex to vertex correspondence is known, along with the seven variants on this metric that involve geometrical transformations; a Monte Carlo approach to adjust transmission error scores for chance; and means for visualizing emergent structure through triangle prototyping and the use of scatterplots in triangle shape-size space. We have also demonstrated the use of a speech synthesizer in training, which offers several benefits over the sole use of the written modality. These are the secondary contributions that this dissertation has made.

4.4. Criticisms

Since the meaning space is unstructured, it was not possible to determine a single metric that could account for the various types of structure that might emerge. This is major problem with this type of experiment, and future work should seek novel and better-founded methods for measuring dissimilarity in an infinite meaning space[34]. We tried to avoid this problem by keeping an open-mind about the type of structure that might emerge and by scrutinizing the languages qualitatively. In fact, there was a high degree of correlation between the kind of structure predicted by the scores and what we observed. However, this may be due to the fact that we used the quantitative results to guide our qualitative exploration of the data. One way to expand on the structure score we have

[34] For example, there is a body of literature that attempts to tackle some of the problems we face (see Alt, Behrends, & Blömer, 1995; Bronstein, Bronstein, & Kimmel, 2006; Veltkamp, 2001), although the methods are general and would need to be customized for our purposes.

developed here would be to consider the reflection property and also consider all properties independently of shape. A better approach would be to run a separate triangle similarity judgement task and use a technique such as multidimensional scaling to determine the dimensions of the underlying phenomenal space.

Cornish (2011, p. 90) makes a good case for using an alternative to the Levenshtein distance when dealing with spoken signals. By using a speech synthesizer, we impose phonological structure on the strings, so this experiment is arguably better-suited to a string-edit metic that is phonetically aware (see Kessler, 2005 for some discussion).

4.5. Conclusion

This project has implemented an infinite meaning space in a human iterated learning experiment. As far as we are aware, this has not been attempted previously. The results showed that iterated learning can give rise to categorical structure even when the space is infinite in magnitude. Separate chains divided the space in subtly different, lineage-specific ways, but, in general, participants showed a strong bias for the size and shape properties of our stimuli. We have also noted an effect of sound symbolism.

We hope that this experiment will usher in a movement away from fixed, monolithic meaning spaces to ones that are more representative of our world. Future work could go in two directions: (1) additional work on the paradigm we introduce here, or (2) more complex spaces tending ever closer to the infinity of our universe.

Appendices

A. Recruitment invitation

The following advertisement was used to recruit participants. The text was posted on the University of Edinburgh's SAGE recruitment website.

> **Earn £5.50 for taking part in a language learning experiment**
> We are seeking native speakers of English for a language learning task. The language you will be learning is from another universe – the curious two-dimensional world of the Flatlanders. The task should take no longer than 45 minutes, and you will be paid £5.50 on completion. The participant who learns the language best will also win a £20 Amazon voucher! Please sign up for a slot using this link: http://doodle.com/26p6vxk8xn8z2chc Please email me once you've signed up for a slot to confirm your appointment. We will meet in the foyer of the Dugald Stewart Building on the University of Edinburgh campus. If you need to cancel or change your appointment, please let me know as soon as possible. Any other questions, feel free to email me at j.w.carr@sms.ed.ac.uk

B. Written briefing

Prior to taking part in the experiments, participants read the following text. Underlined text applies to experiment 2 only.

> You have just entered a parallel universe that has only two dimensions! This curious place is inhabited by an intelligent life form, the Flatlanders, who are obsessed with two-dimensional shapes and have a huge vocabulary just for triangles alone.
>
> Your task is to learn the words that the Flatlanders use for triangles to help us establish contact with these strange beings. It's a pretty difficult task — but we think you're the right person for the job!
>
> **Stage 1: Training**
> You will see a series of triangles, one by one. Each triangle will be presented with its name in the Flatlander language. The name will also be pronounced by the computer to help you learn it. After every third triangle, you will see one of those three triangles again and you must type in its name. This stage is designed to help you learn the language.
>
> **Stage 2: Test**
> Again, you will see a series of triangles. For each triangle, simply type in what you think it's called based on the training you completed in stage 1. <u>However, if you use the same word too frequently, you will see a message asking you to use a different word.</u> The test is designed to assess how well you've learned the Flatlander language, and there's a £20 Amazon voucher for whoever learns it best.
>
> You will learn a lot of words very quickly, and it may be difficult to take it all in. But don't panic! The most important thing is to maintain good relations with the Flatlanders by

giving it your best shot. You must type in an answer for every triangle, but it's okay to guess if you're unsure. Even if you get the word wrong, you'll still get points for getting the word partially correct.

Good luck!

C. Oral briefing

The text below is an approximate phrasing of the oral briefing that participants received before starting the experiment. Underlined text applies to experiment 2 only.

There are two stages, the training and the test. In the training, you'll see three triangles, one at a time, each presented with the word to describe it, and then you'll see one of those triangles again and you just have to type in what the word was for that triangle. And then this repeats a whole bunch of times. Then in the test, you'll see a series of triangles, again one at a time, and for each one you just type in what you think the word is based on the training you did in the first stage. When you're going through the test, you should try to do each one in less than ten seconds on average, otherwise it'll just take you ages. Also, if you try to use the same word lots of times, you'll see a little message asking you to use another word – it's just to stop you repeating the same word too frequently.

D. Post-test questionnaire

Immediately after completing the experiment, participants answered the following questions. Underlined question applies to experiment 2 only.

1. How difficult was it to learn and recall the words?

very easy **very hard**

1 2 3 4 5 6 7 8 9 10

2. Did you identify any patterns in the triangle images? You can draw in the boxes below to help explain any patterns you noticed.

3. Did you identify any patterns in the words?

4. What strategies did you use (if any) to learn the words?

5. The triangles presented in the test phase were not the same as those you learned in training. Did you notice this?

6. Did you see the message that warned you against using the same word too frequently? Did this restrict you from using a word you thought was correct?

7. Do you have any experience learning languages, or do you speak any languages other than English?

E. Debriefing

Participants were given the following debrief information to take away with them.

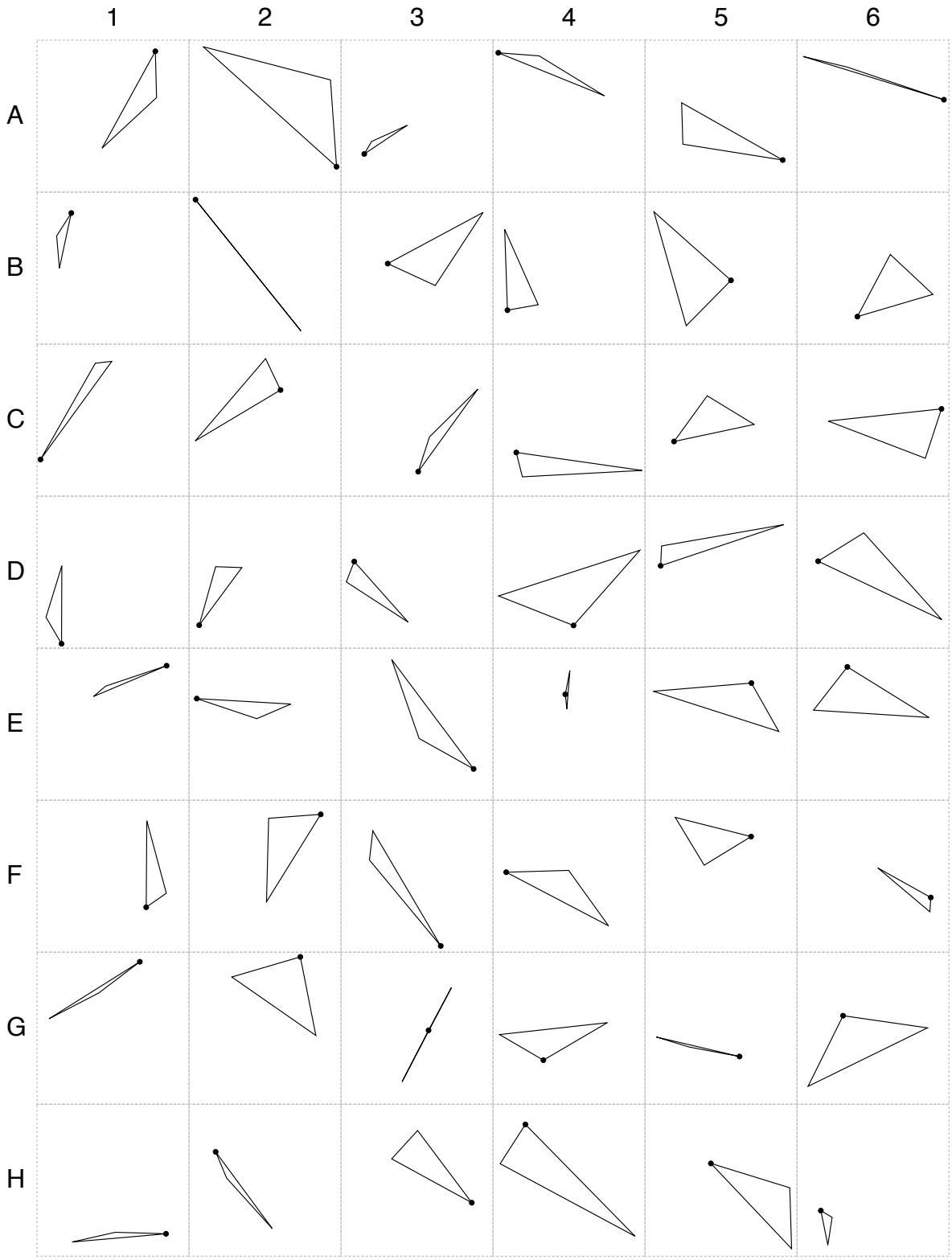
> This experiment is one of many that we are conducting at Edinburgh University to help us understand how languages are learned and how they evolve. The training that you completed was based on the answers given by another participant. In turn, the answers that you gave will become the training material for a future participant. This allows us to create chains of learners and look at how the language changes as it is passed from one person to another.
>
> In this experiment, we wanted to see how participants would respond to an "infinite meaning space". The meaning space in this experiment refers to all the possible triangles that exist in the Flatlanders' world. It is infinite because the triangles were generated randomly such that there are essentially an infinite number of possible triangles. This reflects the infinite number of meanings that could be expressed in our own world using natural languages.
>
> The artificial languages start out random and chaotic, with no systematic way of describing the possible shapes. After several generations, however, we predict that the languages will begin to evolve structure, so that certain words or certain parts of the words will begin to map on to particular aspects of the shapes. For example, a system might emerge where the first syllable describes which way the triangle is pointing, the second syllable describes the thickness of the the triangle, and the third syllable describes its location on the screen.
>
> To motivate you to learn the words as best as you could, we told you that there was a prize for the best learner. In fact, the prize winner will be selected at random. This is because we have no way of objectively measuring the performance of individuals given that the languages are in a constant state of flux and some will be easier to learn than others. We also didn't tell you beforehand that the items you would be tested on were different from those you were trained on. We did this to see whether participants could generalize what they learned in training to the test component of the experiment.
>
> We take our ethical responsibilities seriously, and we hope you do not feel too deceived. Your information and results will be stored safely and will remain confidential. Personally identifiable information will never be linked to your results or published anywhere. However, if you have concerns about how your information will be used or you would like to withdraw your results from the study, please contact Jon Carr at j.w.carr@sms.ed.ac.uk

F. Data provenance

All code has been published and versioned on GitHub: http://github.com/jwcarr/infinity. The experiment and analyses were run on a MacBook Air running OS X Lion. Version numbers of software, libraries, dependencies, and modules used to run the experiment and analyses: Apache 2.2.22, GCC 4.2.1, Google Chrome 27.0, matplotlib 1.2.1, NumPy 1.5.1, pcor 1.0, PHP 5.3.15, Python 2.7.4, python-Levenshtein 0.10.2, R 3.0.1, randomdotorg 0.1.3a1, SciPy 0.12.0. Raw data files can be downloaded from https://www.dropbox.com/s/tqbw7hmlngblt88/Flatlanders_2013.zip

G. Stimuli in the STABLE SET

All 48 stimuli in the STABLE SET are presented below in no particular order. A coordinate system is provided to aid referencing of the stimuli.



References

Abbott, E. A. (1884). *Flatland: A romance of many dimensions*. London, UK: Seeley & Co.

Ahlner, F., & Zlatev, J. (2010). Cross-modal iconicity: A cognitive semiotic approach to sound symbolism. *Sign Systems Studies*, *38*, 298–347.

Alt, H., Behrends, B., & Blömer, J. (1995). Approximate matching of polygonal shapes. *Annals of Mathematics and Artificial Intelligence*, *13*, 251–265. doi:10.1007/BF01530830

Apple Computer. (2006). Speech synthesis programming guide. Retrieved from https://developer.apple.com/library/mac/documentation/userexperience/Conceptual/SpeechSynthesisProgrammingGuide/SpeechSynthesisProgrammingGuide.pdf

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge, UK: Cambridge University Press.

Beale, J. M., & Keil, F. C. (1995). Categorical effects in the perception of faces. *Cognition*, *57*, 217–239. doi:10.1016/0010-0277(95)00669-X

Bickerton, D. (2007). Language evolution: A brief guide for linguists. *Lingua*, *117*, 510–526. doi:10.1016/j.lingua.2005.02.006

Bronstein, A. M., Bronstein, M. M., & Kimmel, R. (2006). Efficient computation of isometry-invariant distances between surfaces. *SIAM Journal on Scientific Computing*, *28*, 1812–1836. doi:10.1137/050639296

Brown, J. E. (2008). Literacy, linguistics and compositionality: Investigating the effects of cultural systems on learning and language. (Unpublished master's dissertation). University of Edinburgh, Edinburgh, UK.

Caldwell, C. A., & Millen, A. E. (2008). Studying cumulative cultural evolution in the laboratory. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*, 3529–3539. doi:10.1098/rstb.2008.0133

Carr, J. W. (2012). The effects of modified variables on an iterated learning model of linguistic evolution by cultural transmission. In T. C. Scott-Phillips, M. Tamariz, E. Cartmill, & J. R. Hurford (Eds.), *The evolution of language: Proceedings of the 9th international conference* (pp. 416–417). Singapore: World Scientific. doi:10.1142/9789814401500_0058

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*, 489–558. doi:10.1017/S0140525X08004998

Cornish, H. (2011). Language adapts: Exploring the cultural dynamics of iterated learning. (Unpublished doctoral thesis). University of Edinburgh, Edinburgh, UK.

Croft, W. (2000). *Explaining language change: An evolutionary approach*. London, UK: Longman.

Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London, UK: John Murray.

del Giudice, A. (2012). The emergence of duality of patterning through iterated learning: Precursors to phonology in a visual lexicon. *Language and Cognition*, *4*, 381–418. doi: 10.1515/langcog-2012-0020

Duda, R. O., Heart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). New York, NY: Wiley.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*, 429–492. doi: 10.1017/S0140525X0999094X

Fay, N., Garrod, S., Roberts, L., & Swoboda, N. (2010). The interactive evolution of human communication systems. *Cognitive Science*, *34*, 351–386. doi:10.1111/j. 1551-6709.2009.01090.x

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science*, *29*, 737–767. doi:10.1207/s15516709cog0000_34

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. Cambridge, MA: MIT Press.

Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of representation: Where might graphical symbol systems come from? *Cognitive Science*, *31*, 961–987. doi:10.1080/03640210701703659

Goldstone, R. L., & Hendrickson, A. T. (2009). Categorical perception. *WIREs Cognitive Science*, *1*, 69–78. doi:10.1002/wcs.26

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, *31*, 441–480. doi:10.1080/15326900701326576

Harnad, S. (1987). *Categorical perception: The groundwork of cognition*. Cambridge, UK: Cambridge University Press.

Horner, V., Whiten, A., Flynn, E., & de Waal, F. B. M. (2006). Faithful replication of foraging techniques along cultural transmission chains by chimpanzees and children. *Proceedings of the National Academy of Sciences of the USA*, *103*, 13878–13883. doi:10.1073/ pnas.0606015103

Hurford, J. R. (1989). Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua*, *77*, 187–222. doi:10.1016/0024-3841(89)90015-6

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford, UK: Oxford University Press.

Keller, R. (1994). *On language change: The invisible hand in language*. (B. Nerlich, Trans.). London, UK: Routledge.

Kessler, B. (2005). Phonetic comparison algorithms. *Transactions of the Philological Society*, *103*,

243–260. doi:10.1111/j.1467-968X.2005.00153.x

King, R. (2011). Exploring expressivity: A closer look at the evolution of linguistic structure. (Unpublished master's dissertation). University of Edinburgh, Edinburgh, UK.

Kirby, S. (1999). *Function, selection, and innateness: The emergence of language universals*. Oxford, UK: Oxford University Press.

Kirby, S. (2002a). Learning, bottlenecks and the evolution of recursive syntax. In E. J. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models* (pp. 173–203). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511486524.006

Kirby, S. (2002b). Natural language from artificial life. *Artificial Life*, *8*, 185–215. doi:10.1162/106454602320184248

Kirby, S. (2007). The evolution of meaning-space structure through iterated learning. In C. Lyon, C. L. Nehaniv, & A. Cangelosi (Eds.), *Emergence of communication and language* (pp. 253–267). London, UK: Springer-Verlag. doi:10.1007/978-1-84628-779-4_13

Kirby, S., & Hurford, J. R. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi & D. Parisi (Eds.), *Simulating the evolution of language* (pp. 121–147). London, UK: Springer Verlag.

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 of the USA*, *105*, 10681–10686. doi:10.1073/pnas.0707835105

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences of the USA*, *104*, 5241–5245. doi:10.1073/pnas.0608222104

Laland, K. N., & Williams, K. (1997). Shoaling generates social learning of foraging information in guppies. *Animal Behaviour*, *53*, 1161–1169. doi:10.1006/anbe.1996.0318

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*, 299–321. doi:10.1016/0885-2014(88)90014-7

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*, 707–710.

Liberman, A. M., Harris, K. S., Hoffman, H. S. & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*, 358–368. doi:10.1037/h0044417

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PLoS ONE*, *5*, e8559. doi:10.1371/journal.pone.0008559

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, *27*, 209–220.

Matthews, C. (2009). The emergence of categorization: Language transmission in an

iterated learning model using a continuous meaning space. (Unpublished master's dissertation). University of Edinburgh, Edinburgh, UK.

Maynard Smith, J., & Szathmáry, E. (1995). *The major transitions in evolution*. Oxford, UK: Oxford University Press.

Mesoudi, A. (2011). *Cultural evolution: How Darwinian theory can explain human culture and synthesize the social sciences*. Chicago, IL: University of Chicago Press.

Mesoudi, A., & Whiten, A. (2008). The multiple roles of cultural transmission experiments in understanding human cultural evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*, 3489–3501. doi:10.1098/rstb.2008.0129

Monaghan, P., Christiansen, M. H., & Fitneva, S. A. (2011). The arbitrariness of the sign: Learning advantages from the structure of the vocabulary. *Journal of Experimental Psychology: General*, *140*, 325–347. doi:10.1037/a0022924

Murray, K. (2009). Issues of literacy, issues of modality: Language evolution from a cultural perspective. (Unpublished master's dissertation). University of Edinburgh, Edinburgh, UK.

Murray, L. (2010). Iterated learning with human subjects: Adding communication and feedback. (Unpublished master's dissertation). University of Edinburgh, Edinburgh, UK.

Nuckolls, J. B. (1999). The case for sound symbolism. *Annual Review of Anthropology*, *28*, 225–252. doi:10.1146/annurev.anthro.28.1.225

Nygaard, L. C., Cook, A. E., & Namy, L. L. (2009). Sound to meaning correspondences facilitate word learning. *Cognition*, *112*, 181–186. doi:10.1016/j.cognition.2009.04.001

Oliphant, M. (1996). The dilemma of Saussurean communication. *BioSystems*, *37*, 31–38.

Ó Searcóid, M. (2007). *Metric spaces*. London, UK: Springer-Verlag.

Page, E. (1963). Ordered hypotheses for multiple treatments: A significance test for linear ranks. *Journal of the American Statistical Association*, *58*, 216–230.

Parault, S., & Schwanenflugel, P. (2006). Sound-symbolism: A piece in the puzzle of word learning. *Journal of Psycholinguistic Research*, *35*, 329–351. doi:10.1007/s10936-006-9018-7

Perfors, A., & Navarro, D. (2011). Language evolution is shaped by the structure of the world: An iterated learning analysis. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd annual conference of the Cognitive Science Society* (pp. 477–482). Austin, TX: Cognitive Science Society.

Pinker, S. (1994). *The language instinct*. London, UK: Penguin.

Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, *13*, 707–784. doi:10.1017/S0140525X00081061

Pullum, G., & Scholz, B. (2002). Empirical assessment of stimulus poverty arguments. *The Linguistic Review*, *19*, 9–50. doi:10.1515/tlir.19.1-2.9

Richerson, P. J., & Boyd, R. (2005). *Not by genes alone: How culture transformed human evolution*. Chicago, IL: University of Chicago Press.

Roberson, D., Davidoff, J. B., Shapiro, L. (2002). Squaring the circle: The cultural relativity of 'good' shape. *Journal of Cognition and Culture*, *2*, 29–51. doi:10.1163/15685370 2753693299

Schoenemann, P. T. (2009). Evolution of brain and language. *Language Learning*, *59*, 162–186. doi:10.1111/j.1467-9922.2009.00539.x

Scott-Phillips, T. C., & Kirby, S. (2010). Language evolution in the laboratory. *Trends in Cognitive Sciences*, *14*, 411–417. doi:10.1016/j.tics.2010.06.006

Scott-Phillips, T. C., Kirby, S., & Ritchie, G. R. S. (2009). Signalling signalhood and the emergence of communication. *Cognition*, *113*, 226–233. doi:10.1016/j.cognition. 2009.08.009

Selten, R., & Warglien, M. (2007). The emergence of simple languages in an experimental coordination game. *Proceedings of the National Academy of Sciences of the USA*, *104*, 7361–7366. doi:10.1073/pnas.0702077104

Seyfarth, S. (2010). Auditory diffusion chains as a laboratory method for studying sound change and cultural evolution. (Unpublished undergraduate dissertation). Northwestern University, Evanston, IL, USA.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323. doi:10.1126/science.3629243

Smith, K. (2004). The evolution of vocabulary. *Journal of Theoretical Biology*, *228*, 127–142. doi:10.1016/j.jtbi.2003.12.016

Smith, K., Brighton, H., & Kirby, S. (2003). Complex systems in language evolution: The cultural emergence of compositional structure. *Advances in Complex Systems*, *6*, 537–558. doi:10.1142/S0219525903001055

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *116*, 444–449. doi:10.1016/j.cognition.2010.06.004

Steels, L. (1997). The synthetic modeling of language origins. *Evolution of Communication*, *1*, 1–34. doi:10.1075/eoc.1.1.02ste

Studdert-Kennedy, M. (2005). How did language go discrete? In M. Tallerman (Ed.) *Language origins: Perspectives on evolution* (pp. 48–67). Oxford, UK: Oxford University Press.

Veltkamp, R. C. (2001). Shape matching: Similarity measures and algorithms. In *SMI 2001 Conference on Shape Modeling and Applications* (pp. 188–197). IEEE Computer Society. doi: 10.1109/SMA.2001.923389

Verhoef, T. (2012). The origins of duality of patterning in artificial whistled languages. *Language and Cognition*, *4*, 357–380. doi:10.1515/langcog-2012-0019

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007).

Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences of the USA*, *104*, 7780–7785. doi:10.1073/pnas.0701644104

Winters, J. (2009). Adaptive structure, cultural transmission and language: Investigating a population dynamic in human iterated learning. (Unpublished master's dissertation). University of Edinburgh, Edinburgh, UK.

Zuidema, W., & de Boer, B. (2009). The evolution of combinatorial phonology. *Journal of Phonetics*, *37*, 125–144. doi:10.1016/j.wocn.2008.10.003