Iconicity emerges and is maintained in spoken language

David Vinson[a], Matthew Jones[a], David M. Sidhu[a], Alex Lau-Zhu[b], Julio Santiago[c] & Gabriella Vigliocco[a,1]

[a]Division of Psychology and Language Sciences, University College London, WC1E 6BT, UK

[b]Kings College London, WC2R 2LS, UK

[c]Dept. de Psicología Experimental y Fisiología del Comportamiento, Facultad de Psicología, Universidad de Granada, Campus de Cartuja s/n 180179-Granada, SPAIN

[1]Corresponding author: g.vigliocco@ucl.ac.uk

Word count: 9386

Keywords: *Iconicity, Sound-symbolism, Cultural Evolution, Iterated Learning, Language Production*

Accepted for publication in ***Journal of Experimental Psychology: General***

Abstract

Iconicity is the property whereby signs (vocal or manual) resemble their referents. Iconic signs are easy to relate to the world, facilitating learning, and processing. Here we examine whether the benefits of iconicity will lead to its emergence and to maintenance in language. We focus on shape iconicity (the association between rounded objects and round-sounding words like *bouba*; and between spiky objects and spiky-sounding words like *kiki*) and motion iconicity (the association between longer words and longer events). In Experiment 1 participants generated novel labels for round vs. spiky shapes, and long vs. short movements (1a: text, 1b: speech). Labels for each kind of stimulus differed in a way that was consistent with previous studies of iconicity. This suggests that iconicity emerges even on a completely unconstrainted task. In Experiment 2 (2a: text, 2b: speech), we simulate language change in the laboratory (as iterated learning) and show that both forms of iconicity are introduced and maintained through generations of language users. Thus, we demonstrate emergence of iconicity in spoken languages and we argue that these results reflect a pressure for language systems to be referential which favours iconic forms in the cultural evolution of language (at least up to a point where it is balanced by other pressures, e.g., discriminability). This can explain why we have iconicity across natural languages and may have implications for debates on language origins.

Introduction

Iconicity is the property of signs (including words, gestures, and signs in signed languages) whereby they resemble their referents. Although iconicity has long been recognized to be present in languages (both spoken and signed, Perniss, Thompson & Vigliocco, 2010), because it is not as clearly visible in Indo-European languages it has been largely neglected. Instead, arbitrariness, rather than iconicity, has been taken as the hallmark of languages (Locke, 1690; De Saussure, 1983/1916). This is the position according to which there is no special relationship between signs and their meanings (see Dingemanse et al., 2015). However, iconicity is far from negligible across spoken and signed languages, and more recent proposals consider iconicity and arbitrariness to be complementary forces in language (see Dingemanse et al., 2105; Perniss & Vigliocco, 2014). In this study, we address the key questions of whether iconicity emerges when speakers create new labels, and whether they emerge and are maintained through generations of speakers.

Types of Iconic Mappings

The most prominent example of iconicity in spoken language is *onomatopoeia*: instances in which the phonology of a word directly imitates sounds in the real-world (e.g., *buzz*, *bang*). Onomatopoeia have consistently been found in all spoken languages (Perniss et al., 2010). However, iconicity can also be present through *sound symbolic associations*: associations between certain language sounds and properties. The most well-known example of this is the maluma/takete effect (i.e., shape sound symbolism; Köhler, 1929): an association between voiced stops (e.g., /b, d, g/), sonorant consonants (e.g., /l, m, n/) and back rounded vowels (e.g., /ɑ/ as in *hot*), with round shapes; and voiceless stop consonants (e.g., /p, t, k/) and front unrounded vowels (e.g., /i/ as in *heat*) with spiky shapes (e.g., McCormick et al., 2015; see Sidhu & Pexman, 2018 for a discussion of mechanisms underlying these

associations). In this way a word like *balloon* could be considered iconic in that its phonemes evoke associations resembling its meaning.

Shape sound symbolism has been the most well studied example of sound symbolism (for reviews see Lockwood & Dingemanse, 2015; Sidhu & Pexman, 2018). A plethora of studies have demonstrated that when participants are asked to pair nonword labels with either round or sharp shapes, they do so in a manner that is consistent with shape sound symbolism (e.g., Nielsen & Rendall, 2012; Sidhu & Pexman, 2016). However, the majority of these studies have involved experimental conditions that may accentuate the iconic bias. For instance, participants are most often presented with a pair of nonwords (e.g., *bouba* and *kiki*) which may highlight the relevant phonological contrast. Similarly, they are most often presented with a pair of shapes (i.e., a round and a jagged one) which might highlight the relevant visual contrast. Indeed, Ahlner and Zlatev (2010) speculated that these two contrasts are necessary for establishing the iconic link between nonwords and shapes. Even when this pairing is not present in a single trial (e.g., asking participants to rate the fit between a single nonword and shape; Cuskley, Simner, & Kirby, 2015), it is implied over the course of multiple trials. Thus, the existing literature leaves open the question of whether shape iconicity will emerge and affect the creation of labels under less constrained, more naturalistic, conditions. Investigating this was one of the main goals of the present study.

Iconicity is also expressed beyond the level of individual phonemes, at the supra-segmental level. For example, language surveys have noted instances in which reduplication of syllables indicates repetition of action or events (e.g., the Japanese mimetic word *goro* means “a heavy object rolling”, while *gorogoro* means “a heavy object rolling repeatedly”; Vigliocco & Kita, 2006). Another example is when vowels are lengthened to indicate that the duration of an event that extends in time (e.g., the Siwu word *dzoro* means “long”, while *dzoroo* means “very long”; Dingemanse et al., 2015; Dingemanse et al., 2015). These

observations seem to suggest a general association between word length and event duration. However, this has not been thoroughly investigated experimentally. Perlman et al. (2015) demonstrated that participants generate longer vocalizations as clues for the meaning *long* (as compared to *short*). However, this study included the words "long" and "short" as targets, rather than observable events of different durations. In addition, participants produced *non-linguistic* vocalizations.

A second goal of this paper was to investigate the potential association between word length and event duration, using observable events as targets, and having participants generate nonwords rather than non-linguistic vocalizations. There is evidence of such an association for the closely related dimension of speed. Cuskley (2013) found that when asked to adjust the speed of an object to match a given nonword, back vowels were associated with slower movements, and consonant reduplication was associated with faster movements. We examined whether an association between word length and event duration would affect the creation of labels for events lasting different amounts of time.

The Emergence and Maintenance of Iconicity

Even if shape and motion iconicity are shown to emerge when speaker create new labels, it is not guaranteed that they will be *maintained* in a lexicon over time. That is, even if shape and motion iconicity affect the creation of labels, it is not guaranteed that this iconicity will survive and proliferate in a real language over generations of speakers. Examining this was the third and final goal of the present study. There are principled reasons to expect iconicity to survive in a language if it provides some benefit to speakers. This is based on work that makes analogies between cultural evolution (e.g. language change) and natural selection (Christiansen & Chater, 2008; Kirby, Smith, & Brighton, 2004). Essentially, linguistic forms or rules that are easier to learn, produce and/or process outcompete alternatives by spreading faster and being used more often.

Indeed, research has shown that iconic words are easier to learn (for reviews see Imai & Kita, 2014; Perniss & Vigliocco, 2014; Nielsen & Dingemanse, 2020). For example, the *Sound-Symbolism Bootstrapping Hypothesis* claims that, among other things, pre-verbal infants are sensitive to sound symbolism, due to a biologically endowed ability to map and integrate multi-modal input, and that this helps them associate speech sounds with their referents (Imai & Kita, 2014; see also Perniss & Vigliocco, 2014 for a similar proposal). There is evidence that more iconic signs/words are learnt first (Caselli & Pyers, 2017; Laing, 2014; Perry et al., 2017; Thompson et al., 2012). Moreover, a number of studies have found that more iconic words are easier to learn, even beyond infancy. Imai, Kita, Nagumo, and Okada (2008) found that Japanese-speaking two- and three-year old children were better able to learn verbs for manner of walking if they were iconic vs. when they were not. This was later replicated with English-speaking children (Kantartzis, Imai, & Kita, 2011). Other studies have also found an iconicity benefit with adults. For example, Lockwood, Dingemanse, and Hagoort (2016) found that Dutch speakers were better able to remember Japanese mimetic words when they were learned along with their correct definitions as opposed to the opposite (see also Lockwood, Hagoort, & Dingemanse, 2016).

In addition to improved learnability, there is also evidence of iconic words being easier to process. Recent work demonstrates advantages for onomatopoeia in processing by English-speaking aphasic patients (Meteyard, Stoppard, Snudden, Cappa, & Vigliocco, 2015). Sidhu, Vigliocco and Pexman (2020) also showed a benefit for onomatopoeic words in a healthy population. Finally, there is also evidence from neuroimaging studies that iconic words may elicit richer representations during processing (e.g., Kanero, Imai, Okuda, Okada, & Matsuda, 2014; Vigliocco, Zhang, Del Maschio, Todd & Tuomainen, 2020). Learnability and processing advantages of iconicity provide a principled reason to expect it to be maintained in language.

It is important to note that we would not expect iconicity to increase in a language indefinitely. This is because arbitrariness also conveys advantages onto language, and thus will also increase. Namely, arbitrariness aids discriminability (Hockett, 1960; Monaghan, Christiansen, & Fitneva, 2011). This is because when a word is not required to resemble its meaning, it is free to adopt any form, leading to greater variety (and thus discriminability) among words. Conversely, in an iconic vocabulary, similar meanings would beget similar forms, resulting in confusion (see Monaghan et al., 2011; Sidhu & Pexman, 2018). Thus, we would expect iconicity to increase up to a point, but to be balanced out by the need for words in a language to be discriminable (and thus non-iconic).

Iterated Language Learning as a Simulation of Cultural Evolution

In order to explore the emergence of subsequent maintenance of iconicity in language, we employ an iterated learning paradigm (for reviews see Kirby et al., 2014; 2015). In a typical iterated learning study, a participant learns a set of words and then has to reproduce them during a testing round. When reproducing the learned words at test, participants will inevitably make errors (e.g., misremembering *pilu* as *kinepilu*). Their reproductions (including the mistakes) become the learning set for the next participant (e.g., the next participant would learn *kinepilu*). This is repeated several times. This paradigm is based on the fact that language is culturally transmitted and explores the features that can emerge through such iterated learning. It examines how biases affect cultural transmission, and thus the resulting language in an idealized model of language evolution. The model is clearly idealized because any transmission in the lab is necessarily on a vastly smaller scale than during language evolution.

A handful of studies have addressed vocal iconicity in iterated learning. Most of them have focused on prosody (e.g. pitch and voice quality), which can be modulated to iconically

match referents. For example, Perlman et al. (2015) found that individuals could communicate a fixed set of meanings to one another using only highly iconic vocal charades (unword-like vocal sounds). In related work, Verhoef, Roberts, and Dingemanse (2015) showed that iconicity emerged in a model of cultural evolution where participants produced acoustic labels of varying pitch using slide whistles. However, prosody is only one dimension of spoken language, and the potential for iconicity in the wordform (i.e., segmental information) is less explored.

Using an iterated learning approach, Edmiston et al. (2018) demonstrated that a language beginning with imitations of environmental sounds becomes more word-like over time. While this could demonstrate how onomatopoeia plausibly emerged, other studies have focused on an emergence of non-onomatopoeic iconicity. A recent study by Tamariz et al. (2018) explored shape sound symbolism using iterated learning. Their study had two conditions: one in which participants worked in pairs, and another in which they participated individually. Participants learned a language of invented words for round and spiky shapes. The authors found that over generations, words for the round and spiky objects changed to become more iconic. Note that this only occurred in the partner condition. However, Tamariz et al.'s (2018) paradigm did not encourage speakers to create new labels because no additional new items were presented at test to force innovation. Participants were also given feedback on their performance, which emphasizes remembering the given forms. This leaves open the possibility that iconicity might emerge – and be maintained – even when speakers do not interact with a partner.

The Present Study

In the following experiments, we examine the emergence and maintenance of iconicity in language. We investigate two forms of iconicity (i.e., shape and motion), which operate at

two different levels of language (i.e., individual phonemes and suprasegmental features). We begin with a task examining the influence of iconicity on the creation of new labels (Experiment 1) in an unconstrained task. We then explore whether iconicity emerges and is maintained over multiple generations of speakers (Experiment 2) using the iterated learning paradigm. Experiment 2 builds upon Tamariz et al. (2018) in several ways. In addition to presenting words visually, we crucially also examine spoken words. This more accurately approximates the way in which language has been used throughout history. In addition, in order to force participants to innovate, we included images at test for which participants had not learned labels at training (therefore we further did not provide feedback to participants). Finally, and most importantly, we focus on two types of iconicity–shape and motion–which overlap in each stimulus. This increases the complexity of the stimuli, as participants have multiple dimensions which can afford iconicity. This serves to make the task more generalizable, as meanings in the real world often have multiple dimensions that could afford iconicity. It also addresses the concerns mentioned previously, that iconicity tasks often highlight the relevant dimension (e.g., by having stimuli only vary by shape). This provides a challenging test for the maintenance of iconicity in language.

If shape iconicity emerges and persists we would expect labels for round and jagged objects to diverge such that they contain more phonemes previously shown to be associated with roundness and jaggedness, respectively. Similarly, if motion iconicity emerges and persists we would expect labels for longer events to contain a greater number of letters and syllables.

Experiment 1: Do Speakers Create New Iconic Labels?

Previous work has shown that participants show a sound symbolic bias when choosing nonword labels for shapes (e.g., Nielsen & Rendall, 2012), or even constructing a nonword out of available components (e.g., Nielsen & Rendall, 2013). However, these studies have

used contrasting nonwords and shapes, either on an individual trial or over multiple trials. This highlights relevant phonological and visual contrasts, thus leading participants to use strategies that maximise such contrasts.

In Experiment 1 we explored whether iconicity will emerge when participants are asked to generate a nonword for a single referent, without being given options to choose from. In addition, each participant only took part in a single trial, to avoid the overall task context biasing their responses. This experiment also included the first examination of a potential association between word length and event duration.

We presented speakers only two novel stimuli –either rounded or spiky (Figure 1); and a dot moving once or repeatedly. We asked them to make up a new name for it that was not already a real word, and type (Experiment 1a) or speak (Experiment 1b) its name. We then tested how much iconicity was present in these new words. Including writing *and* speech probes the extent to which any bias is unimodal (visual-to-visual) or multimodal (auditory-to-visual).

Methods

Participants

In Experiment 1a, 97 first-year undergraduates at University College London participated as part of a laboratory class. After removing responses that were actual English words and those including a clearly identifiable English word (e.g. “blobby-ku”), 73 remained for the shape analysis (62 women, 52 native English speakers, M_{age} = 19.0, *SD* = 0.8), 71 for the motion analysis (61 women, 48 native English speakers, M_{age} = 19.0, *SD* = 0.8).

Participants in Experiment 1b were 329 visitors to London’s Science Museum, including both children and adults, and speakers of many native languages. Recordings without an audible articulate response were excluded, as were words based on a related

English word. This left 171 participants for the shape task (86 women; 113 native English speakers, M_{age} = 19.4, *SD* = 13.4; Range 5 – 71;), and 194 for the motion task (105 females; 137 native English speakers, M_{age} = 20.3, *SD* = 13.9; Range 4 – 71). The number of participants was not set *a priori*. For both experiments, we had an opportunistic sample made of all the students taking part in a lab class (Experiment 1a); or all visitors to our exhibit at the Science Museum within a two-week period (Experiment 1b)

Materials

For shape iconicity, a set of 16 spiky and 16 rounded shapes was created. The spiky shapes were generated by a Matlab 2012a randomization script based on the procedure reported in Monaghan et al. (2012). The round stimuli were generated by using the GNU image manipulation program (GIMP, 2012) to smooth the spiky shapes' sides into Bezier curves with their fixed points on the spiky shapes' angles, and then matching for size. Stimuli were 600*600 pixel images comprising the shape in black on a white background (see Figure 1). For motion iconicity there were just two silent video stimuli each lasting six seconds. In the single-motion condition, a small black dot made a single upwards stroke (lasting 1.25sec) and then remained still for the remaining 4.75 seconds. For the repeated-motion condition, a dot of the same size moved up and down twice (each movement lasting 1.25sec) and then was still for the final second.

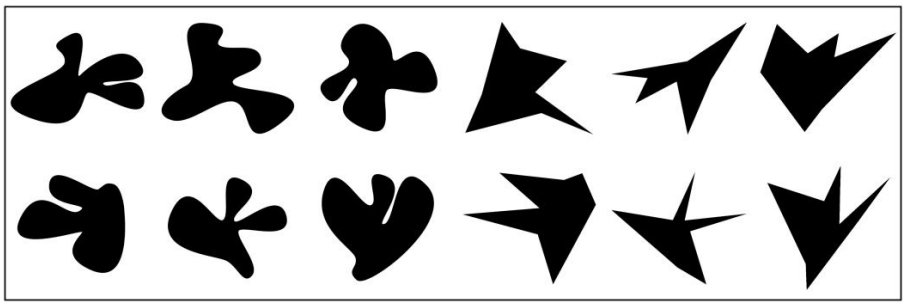

Figure 1: Examples of round and spiky shapes used in Experiment 1 (stimuli for the motion-iconicity condition were a dot moving up and down one vs. four times).

Apparatus and Procedure

Each participant was shown one of the 32 shape iconicity stimuli (randomly selected) and was asked to come up with a name for each stimulus that was not already a real word, and type their response (Experiment 1a) or speak into a microphone (Experiment 1b). They were then shown one of the two motion stimuli, and were asked to provide a name for it as well. All experiments were run using E-Prime 2.0.

Analysis

To analyze sound-shape iconicity in Experiment 1a, we developed a LetterScore index of shape iconicity prior to running the experiment as follows: Monolingual English speakers (N = 28, 12 women, M_{age} = 28.5, *SD* = 12.0 years old) rated 85 consonant-vowel pairings(e.g. 'mu', 'ka'), including all those possible in English orthography with the following exceptions: CV syllables including 'q' and 'x' were left off the list as they are exceptionally rare/absent in English; 'c' and 'g' were excluded as their pronunciation is ambiguous. Participants were asked to make their ratings on the basis of the syllable's sound, using a ten-point scale anchored by a circle (1) representing most round and a star (10) representing most spiky (see Supplementary Materials, Figure S1). We centered the scale by redefining the mean rating (5.04) as zero. Each letter was assigned a LetterScore by calculating the mean ratings of syllables that letter appeared in. Positive scores represent spikiness, negative scores roundness (e.g. z = 1.06, m = -1.46). Excluded letters (c, g, q, x) were assigned LetterScores of 0. Ratings for consonant-vowel pairings and LetterScores appear in Supplementary Tables S1 and S2). A word's LetterScore is the mean of its letters' LetterScores. As LetterScore ratings were normally distributed (Shapiro-Wilk W = 0.978, p = .227) we conducted *t*-tests.

For Experiment 1b we used a different whole-word measure of shape iconicity: WordScore. These were direct ratings of the sound-shape iconicity of the vocabulary produced by the participant, collected after the study was completed. 29 adult monolingual English speakers (7 women, M_{age} = 31.3, *SD* = 11.2) recruited through Prolific.ac were asked to rate edited speech tokens (counterbalanced for half of the participants), using the online platform Gorilla.sc. The scale and counterbalancing procedures were the same as for LetterScore (see Supplementary Materials, Figure S2). As WordScore ratings significantly departed from normality (*W* = 0.977, *p* =.006) we conducted Mann-Whitney U tests.

Motion iconicity was analyzed using number of letters in Experiment 1a and number of syllables in Experiment 1b. Both measures departed from normality (1a letters *W* = 0.928, *p* < .001; 1b syllables *W* = 0.798, *p* < .001) so we conducted Mann-Whitney U tests.

Results

Shape Iconicity

Experiment 1a. LetterScore differed between the rounded (*n* = 33, *M* = -0.331, 95% CI [-0.457, -0.205], *SD* = 0.357) and spiky conditions (*n* = 40, *M* = 0.234, 95% CI [0.096, 0.371], *SD* = 0.431): with *t*(71) = 6.13, *p* < .001; difference = 0.565, 95% CI [0.381, 0.749], Cohen's *d* = 1.41. Both the rounded (*t*(32) = -5.33, *p* < .001, *d* = 0.928) and the spiky (*t*(39) = 3.43, *p* = .001, *d* = 0.543) conditions are significantly different from zero (Figure 2 left panel). Thus rounded shapes tended to be given round names like 'bomo', whereas spiky shapes were given spiky names like 'zorik'.

Experiment 1b. WordScore differed between the rounded (*n* = 85, *M* = -0.176, 95% CI [-0.363, 0.010], *SD* = 0.865) and spiky conditions (*n* = 86, *M* = 0.219, 95% CI [0.051, 0.388], *SD* = 0.786): with *t*(167) = 3.13, *p* = .002; difference = 0.396, 95% CI [0.146, 0.645], Cohen's *d* = 0.479. The spiky condition is significantly different from zero (*t*(85) = 2.59, *p*

= .011, *d* = 0.279), though the rounded condition is only marginally different (*t*(88) = -1.88, *p* = .064, *d* = 0.204). See Figure 2 (right panel).

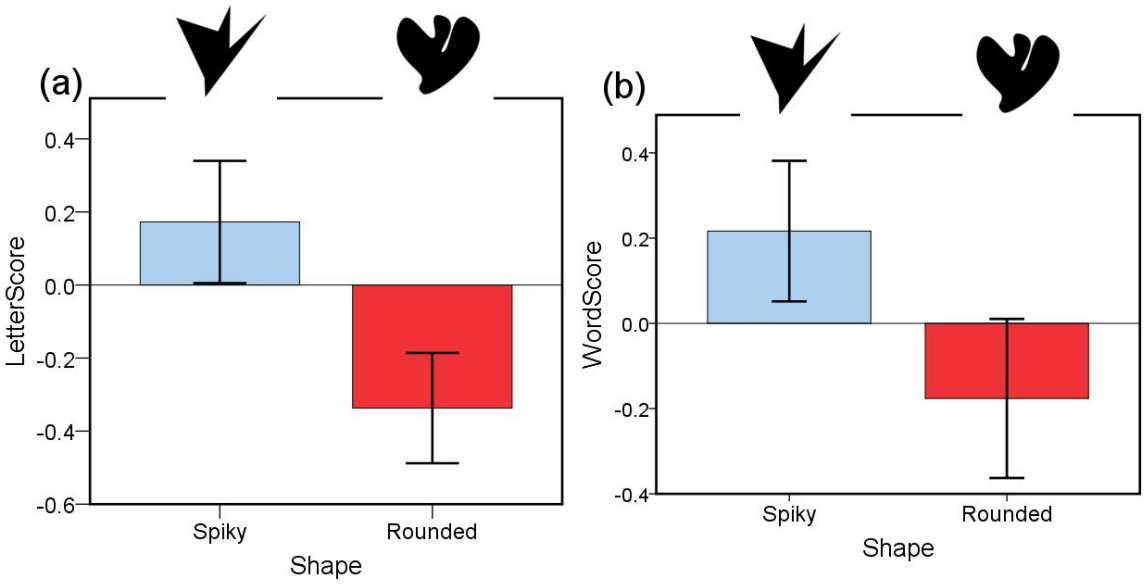

Figure 2: Shape iconicity results from Experiment 1. Positive values = spiky; negative values = rounded. Error bars = 95% CIs. Left panel: Experiment 1a LetterScore; right panel: Experiment 1b WordScore

Motion Iconicity

Experiment 1a. Length differed between the single-motion (*n* = 38, median = 4) and repeated-motion conditions (*n* = 38, median = 5; *W* = 869.5, *p* = .004). See Figure 3 left panel. Thus repeatedly moving stimuli tended to be given long names like ‘lalalalananana’, whereas stimuli that only move a single stroke are given shorter names like ‘lop’.

Experiment 1b. Length did not significantly differ between the single-motion (*n* = 101, median = 2) and repeated-motion conditions (*n* = 100, median = 2; *W* = 5602, *p* = .160); see Figure 3 right panel. This lack of difference may be due to the low sensitivity of number of syllables as a dependent measure (more than 70% of names had only one or two syllables), and may also reflect a sapping of power in this experiment by the noisiness of the museum

environment. But the apparent tendency is similar to that observed under more tightly controlled conditions in Experiment 1a: stimuli that move repeatedly seem to be given long names like ‘tamatandlatu’, whereas stimuli that only move once tend to be given shorter names like ‘bu’.

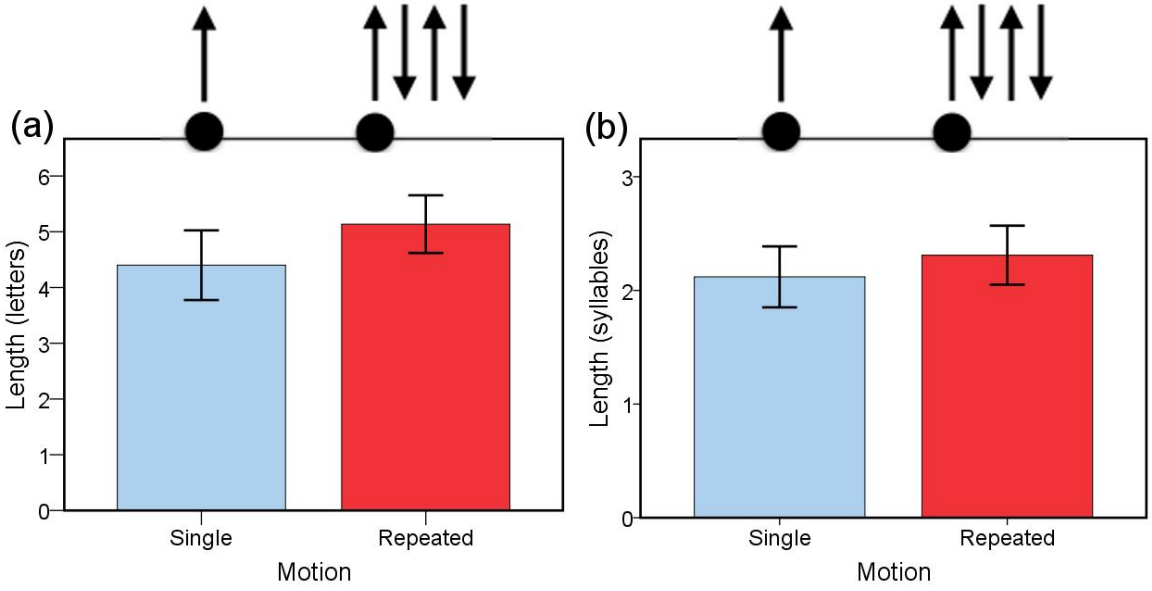

Figure 3: Motion results for Experiment 1. Error bars = 95% CIs. Left panel: Experiment 1a number of letters; right panel: Experiment 1b number of syllables.

Discussion

This experiment demonstrates that iconicity emerges even under very unconstrained task conditions. To our knowledge it is the first demonstration that sound symbolism will affect the generation of labels on a single trial. In addition, we demonstrate the novel effect that nonwords generated for events of a longer duration tend to have longer letter lengths than those for events of a shorter duration. Notably this was only observed in Experiment 1a (with visual stimuli). Perhaps this form of iconicity is more easily supported in visual aspects of language.

We now turn to the emergence *and maintenance* of these forms of iconicity across generations of speakers.

Experiment 2: Do Speakers Change an Arbitrary Language Into a More Iconic One and Is Iconicity Maintained Through Generations of Speakers?

Having demonstrated that participants generate iconic novel labels, we now examine: 1) whether such labels will emerge in a more complex task, and 2) whether they will be maintained across generations. The iterated learning model (Scott-Phillips & Kirby, 2010) has been argued to approximate cultural evolution - specifically language change - using *diffusion chains*, in which successive participants (*generations*) learn from their predecessor. Participant 1 learns a 'language' of novel words for visual stimuli. They are then tested on these names and (unbeknownst to them) names for similar novel stimuli. Testing unseen items compels innovation. Hence when the first generation's responses, including mistakes and changes, are taught to the second, the language evolves. This process is repeated between the second and third generation, and so on. Participants assume this is a simple learning experiment and remain unaware of being in a chain.

If the changes made by participants confer an advantage onto language (e.g., making it more learnable or easier to process), they should be retained across generations (Christiansen & Chater, 2008). If iconicity meets this condition, iconic mappings will emerge from an arbitrary initial language and persist through generations. We would expect novel labels lacking iconicity at first to become more iconic (i.e., words for round vs. spiky referents would diverge such that they became more round- vs. sharp-sounding, respectively; words for longer vs. shorter events should diverge and become longer vs. shorter in length themselves, respectively). Such iconicity, once emerged, would then be maintained across further generations. We test these predictions across ten generations in Experiment 2. In Experiment

2a participants were taught and tested on written words; in Experiment 2b we used spoken words. Specifically, in both Experiment 2a and 2b, we predict an interaction between shape and generation for shape iconicity measures, and an interaction between motion and generation for word length. Moreover, given that this is the first time that more than one iconic dimension is evaluated within a single design, we further carry out exploratory analyses to assess whether shape and motion iconicity interact.

Methods

Participants

Each sub-experiment included 60 participants (Experiment 2a: 32 women, M_{age} = 26.3, *SD* = 8.5 years old; Experiment 2b: 41 women, M_{age} = 22.6, *SD* = 8.2 years old). All participants were native British English speakers from University College London's subject pool. The number of participants was chosen to ensure we did not fall short of the total number of observations in the original Kirby et al (2008) study. As in the original study, the present experiment included 10 generations and 10 participants per chain. However, our focus on shape iconicity implies that we need to directly compare round and spiky shapes, reducing the number of datapoints by participant by one third in comparison to the original study by Kirby et al. (2008) which used three distinct shapes. We therefore increased the number of chains from four to six in the present study to compensate for this.

Some participants were excluded from the study and replaced; they are not included in the demographic summary above. In Experiment 2a, 11 subjects were replaced: five for generating fewer than five word types, two for not being monolingual English speakers, one for noticing testing of unseen items, one for misunderstanding when to type in responses, one for reporting memory problems, and one for having been run on the wrong procedure. In Experiment 2b six participants were replaced: two for generating fewer than five word types,

one for not being a monolingual English speaker, and three because of problems in recording their responses.

Materials

Visual Stimuli. Eighteen 5s video stimuli, varying on dimensions of shape (round/spiky), motion (still/single upwards stroke/up-down bounce) and color (red/green/blue: see Figures 4 and 5 for examples and Supplementary Materials Figure S3 for full set of stimuli). Color was merely used to obtain enough stimuli to make name recall challenging (Kirby, et al., 2008).

Names. We used the LetterScore index of shape iconicity (described in Experiment 1) to create iconically neutral initial vocabularies prior to running Experiment 2, . For the initial language (generation 0), we combined syllables whose LetterScores summed to around zero (e.g. one round = *mo* = -2.22, one spiky = *ti* = 1.74, one neutral = *fa* = 0.06: *motifa*). This ensured that though initial names were neutral, participants had a phonologically varied language. Name length was randomized from between two and four syllables. The initial language was presented in writing in Experiment 2a. In Experiment 2b, the words were recorded by a North American linguist and presented auditorily via headphones (see Supplementary Materials, Table S3).

Unlike Experiment 1a we used both LetterScores and WordScores (see section 2.1.4) to analyze the vocabularies produced by participants in Experiment 2a. The use of both LetterScore and WordScore in Experiment 2a allowed us to overcome potential shortcomings of each one in isolation: LetterScore treating each syllable as equally weighted; WordScore potentially sensitive to context of other words in the vocabulary. Using both measures is also important in this experiment, because unlike Experiment 1a, participants read the labels for SEEN items as well as typing their responses. This might have led to a focus on

orthography/visual processing that might only be reflected in LetterScores and not WordScores. For Experiment 2a WordScores were obtained from 18 participants (nine women, M_{age} = 36.9, *SD* = 11.2 years old) and for Experiment 2b they were obtained from 98 participants (40 women; M_{age} = 32.4, SD = 9.6). All participants were recruited through https://www.prolific.ac and performed the task on https://www.qualtrics.com.

Apparatus and Procedure

Following Kirby, Cornish, and Smith (2008) Experiment 2, participants were assigned to one of six diffusion chains of 10 generations each. They were told they would learn an 'alien language' of word-video pairings. Each language comprised 18 pairings, divided (participant-by-participant) into a SEEN set (12 items) and an UNSEEN set (six items). Participants were trained on the SEEN set only, but (secretly) tested on both sets, to force innovation (see Figure 4 for a schematic of the design). A post-experiment questionnaire confirmed participants typically did not notice the novelty.

Participants learned in three rounds of training, each followed by a testing block. In each round participants were trained on the SEEN set in two randomized orders. In Experiment 2a the first frame of each video was displayed for 1 second before the letter string was displayed below. The video plus name appeared together for 5 seconds. In Experiment 2b, the first frame was also presented for 1 second, followed by the name recording and the rest of the video simultaneously.

In testing, participants saw videos and had to produce their names. In Experiment 2a they typed names using a standard keyboard. In Experiment 2b they spoke names into a microphone. There was no time limit. The first round's testing block contained half the SEEN set and half the UNSEEN set, with the second's containing the other half of each set. The final test featured all items. Responses from the final test became the next generation's

training set. The SEEN set for the next participant was chosen pseudorandomly, with the constraint of minimizing the number of homonyms (i.e. repeated identical words) in order to maximize the available vocabulary size for the next participant in the chain. In Experiment 2b names were edited to remove silence, and volume was normalized.

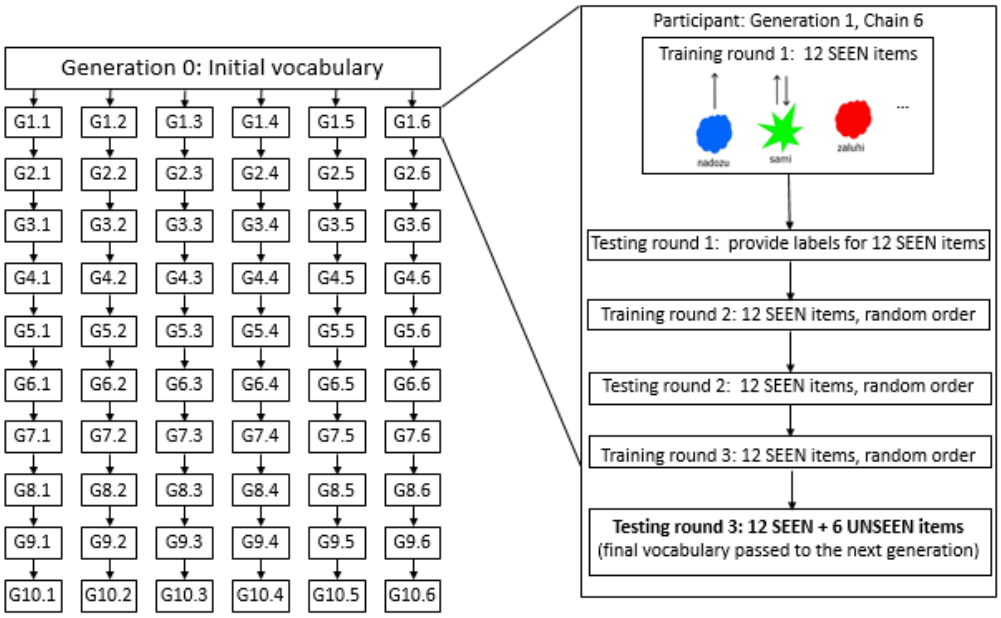

Figure 4: Schematic of experimental design for Experiment 2a/2b. Left side of the figure depicts the way in which participants were assigned to chains and generations. Generation 0: A name was assigned to each item that would be SEEN by participants in generation 1; these were a common starting point for all chains which were otherwise independent of each other. Each rectangle below Generation 0 represents one participant, labeled here by generation and chain number. Names produced by a given participant in generations 1-9 were used to label SEEN items for participants in the next generation. Right side of the figure depicts the sequence of events for an individual participant. SEEN items were labelled based on the previous generation's vocabulary from that chain and were presented in all three training rounds, and appeared in all three testing rounds. UNSEEN items only appeared in testing round 3, and only names from round 3 were used for the next generation. See main text for further details.

Design and Analysis

For statistical analysis, our predictors were generation (0 – 10), shape (round vs. spiky), and motion (still vs. single stroke vs. bounce). In Experiment 2a, shape iconicity was measured by means of LetterScore *and* WordScore; motion iconicity was measured as length

in letters. In Experiment 2b, shape iconicity was measured by WordScore and motion iconicity by length in syllables. Learnability was measured in Experiment 2a as normalised Levenshtein (edit) distance (i.e., error) between instances of the same name between consecutive generations. We only analyzed this measure for Experiment 2a in which responses were typed and thus edit distance on the basis of letters could be straightforwardly calculated.

We employed mixed-effects models using lme4, version 1.1-12 (Bates, Maechler, Bolker, & Walker, 2014) in R version 3.2.3 (R_Core_Team, 2014). *P*-values were from lmerTest version 2.0-29 (Kuznetsova A., Brockhoff, & Christensen, 2016), confidence intervals were calculated using the *t*-distribution, and standard errors and degrees of freedom reported by lme4. Generation was coded linearly and centered. Linear generation terms express an overall directional trend; we also added a quadratic generation term thus permitting us to test for a single bend in this trend. Shape was coded numerically (-0.5 = rounded, +0.5 = spiky), and motion as a factor (still, single stroke, bounce) with still as the reference condition.

Random effects groupings were by generation nested within chain (with the initial generation zero ‘language’ coded as a single chain, to avoid including the same words six times). We aimed for a design-driven maximal random effects structure (Barr, 2013), but were limited in the number of intercepts and slopes we could fit. However, we fit intercepts and slopes for crucial predictors in each particular analysis (shape and its interactions with generation for shape analyses, motion and its interactions with generation for motion analyses, generation and its interactions for all analyses). Where models failed to converge, we removed low variance random effects slopes (with the constraint that main effect slopes of a remaining interaction slope had to be left in place). Analysis scripts and source data can be found at https://osf.io/ysx8t/?view_only=26b67f0426a34bb48994a2a1cbb21117.

For each dependent measure we first fit a hypothesis-driven model, including just the predictor of interest and its interaction with generation (both linear and quadratic generation). For shape iconicity measures (LetterScore and WordScore) this was shape × generation, and for length, this was motion × generation. Although we did not have hypotheses about higher order relationships among these variables, for exploratory purposes we also fit a fully factorial model (shape × motion × generation) for every dependent measure, described below in the Exploratory analyses section.

Results

Shape iconicity

Experiment 2a. Using LetterScore, there were main effects of shape and generation, qualified by significant generation × shape interactions involving both linear and quadratic generation (see Figure 5 for plots, and Table 1 for details of results for shape iconicity). The interaction between shape and linear generation indicated that names for round and spiky shapes diverged in the expected direction over the generations. The significant interaction between shape and quadratic generation further indicated that this divergence slowed in later generations. Using WordScore, again there were main effects of shape and generation, along with significant interactions for both linear and quadratic generation. Names for round and spiky shapes diverged over generations in the predicted direction, slowing in later generations.

Experiment 2b. For WordScore, as in Experiment 2a we found an interaction between generation and shape: names for round and spiky stimuli diverged in the expected direction over generations (see Figure 5, lower right panel).

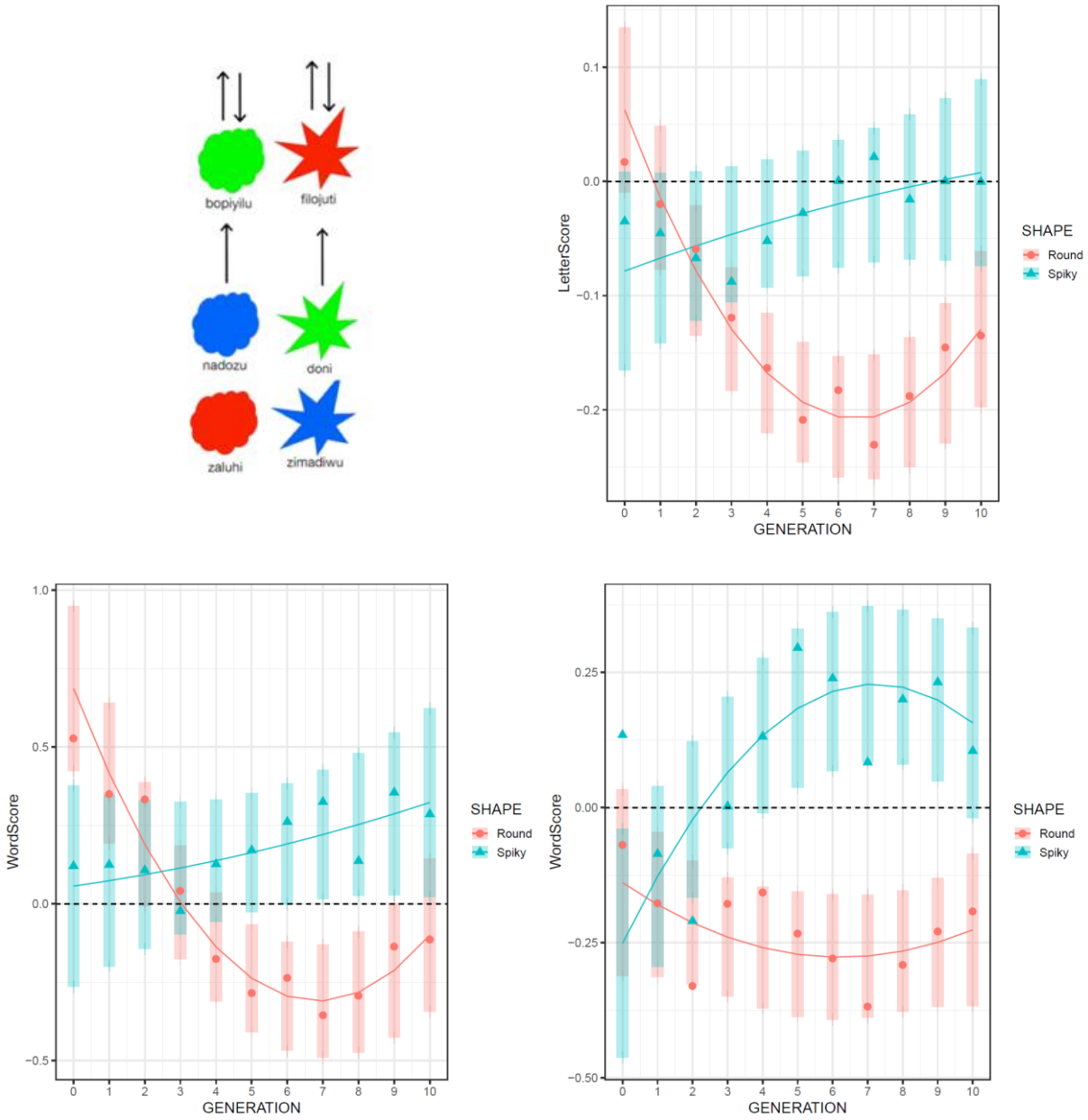

Figure 5: Upper left panel: Examples of items and initial labels for Experiment 2. Arrows indicate the movement of stimuli (repeated bouncing movement; single upward movement; no movement) and were not visible to participants. *Other panels:* Shape iconicity measures as a function of stimulus shape and generation in Experiments 2a and 2b. Higher values indicate more spiky labels and lower values more rounded. Solid lines indicate model fit (linear and quadratic generation) with shaded regions indicating one standard error of model estimate at each value of generation. Solid filled points indicate observed average. *Upper right*: Experiment 2a LetterScore. *Lower left*: Experiment 2a WordScore. *Lower right*: Experiment 2b WordScore.

Table 1. Parameter estimates for shape iconicity measures, Experiments 2a/2b. Reference levels: Generation = 5; Shape = mean(round, spiky).

	Experiment 2a LetterScore					Experiment 2a WordScore					Experiment 2b WordScore				
	Estimate	Std. Error	df	t	Pr(>|t|)	Estimate	Std. Error	df	t	Pr(>|t|)	Estimate	Std. Error	df	t	Pr(>|t|)
(Intercept)	-0.111	0.046	6.63	-2.389	0.050 .	-0.037	0.178	5.91	-0.209	0.841	-0.044	0.097	9.26	-0.454	0.661
Generation(linear)	-0.005	0.010	6.29	-0.554	0.599	-0.026	0.043	6.29	-0.610	0.563	0.016	0.015	55.77	1.061	0.293
Generation(quadratic)	0.003	0.001	48.63	2.766	0.008 **	0.011	0.005	8.97	2.359	0.043 *	-0.003	0.006	56.06	-0.513	0.610
Shape	0.165	0.045	7.51	3.638	0.007 **	0.401	0.116	9.12	3.446	0.007 **	0.455	0.180	7.27	2.531	0.038 *
Generation(linear) × Shape	0.028	0.009	6.39	3.023	0.022 *	0.105	0.023	10.11	4.614	0.001 ***	0.050	0.016	261.12	3.077	0.002 **
Generation(quadratic) × Shape	-0.007	0.002	51.15	-3.436	0.001 **	-0.020	0.007	179.25	-3.011	0.003 **	-0.013	0.006	261.87	-2.165	0.031 *

* p < .05 ** p < .01 *** p < 001

Table 2. Parameter estimates for length measures, Experiments 2a/2b. Reference levels: Generation = 5; Motion = still (no motion).

	Experiment 2a Number of letters					Experiment 2b Number of syllables				
	Estimate	Std. Error	df	t	Pr(>|t|)	Estimate	Std. Error	df	t	Pr(>|t|)
(Intercept)	5.595	0.280	5.37	19.991	0.000 ***	2.520	0.153	6.26	16.420	0.000 ***
Generation	-0.081	0.051	5.74	-1.600	0.163	-0.070	0.022	6.17	-3.141	0.019 *
Shape	-0.391	0.092	1008.48	-4.266	0.000 ***	0.115	0.063	944.34	1.834	0.067
Motion(Bounce)	0.510	0.315	5.55	1.618	0.161	1.236	0.219	6.13	5.637	0.001 **
Motion(Up)	0.372	0.367	5.06	1.016	0.356	0.713	0.197	6.24	3.613	0.010 *
Generation × Shape	0.064	0.031	1008.48	2.048	0.041 *	0.005	0.012	945.03	0.370	0.711
Generation × Motion(Bounce)	0.234	0.078	6.12	3.001	0.023 *	0.138	0.045	6.24	3.090	0.020 *
Generation × Motion(Up)	0.094	0.064	6.02	1.480	0.189	0.157	0.036	6.22	4.363	0.004 **

* p < .05 ** p < .01 *** p < 001

Motion Iconicity

Experiment 2a (Length in Letters). For length, there were no quadratic effects of generation. Crucially, there was an interaction between motion and (linear) generation, particularly driven by a difference between bouncing stimuli and the reference condition (still stimuli; see Figure 6 for plots, and Table 2 for details of results for motion iconicity). An additional analysis excluding still stimuli revealed a significant generation by condition interaction: stimuli with upward movement patterned like still ones (linear generation × motion β = -0.107, SE = 0.046, t = -2.319, p = 0.034). Stimuli that that moved repeatedly acquired long names like ‘piotesque’ whereas still stimuli and those with brief upward movement acquired short names like ‘sami’ (Chain 5, Generation 9).

Experiment 2b (Length in Syllables). As in Experiment 2a there was again an interaction between motion and generation (see Figure 6), but this time both conditions with movement differed from the reference condition (still stimuli): stimuli that move tended to acquire long names like ‘osa-ii-tokai’, whereas still stimuli acquired short names like ‘ofa’ (Chain 2, Generation 8). When still stimuli were excluded from analysis, there was no significant interaction with generation ($|t|$<1.1 for all interactions including generation).

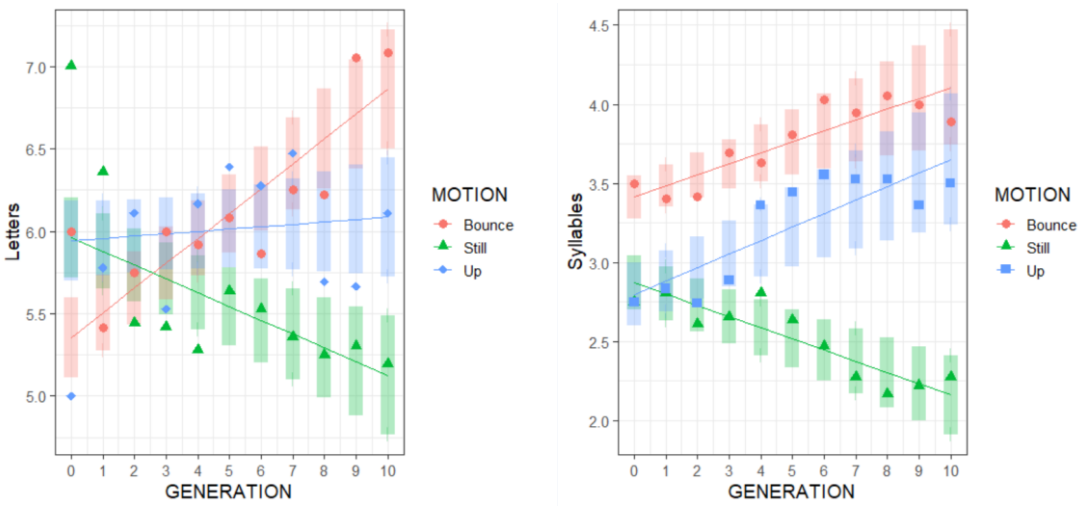

Figure 6. Word length as a function of stimulus motion and generation Experiments 2a and 2b. Solid lines indicate model fit (linear and quadratic generation) with shaded regions indicating one standard error of model estimate at each value of generation. Solid filled points indicate observed average. *Left panel*: Experiment 2a number of letters; *right panel*: Experiment 2b number of syllables.

Error

Error was analyzed for Experiment 2a to test how the languages' learnability changed. Error was operationalized as Levenshtein edit distance between a word and its counterpart in the next generation, normalized to vary between zero and one by dividing by the length of the longer word. We were not able to analyze the spoken words from Experiment 2b in this way, as they did not come ready-coded in discrete symbols easily submitted to such analysis. Because we found effects of both shape and motion iconicity in the analyses reported above, we included both shape × motion and shape × generation interactions in the model. Crucially, there was a sizable main effect of linear generation that did not interact with any other factors: languages became more learnable over time (in keeping with the dynamics of iterated learning). There were also main effects of shape and motion, both qualified by small but reliable interactions with quadratic motion. While all conditions showed the linear decrease in error over generations, there was a greater amount of curvature in this trend for motionless stimuli compared to moving ones, and for spiky stimuli compared to round ones (see Figure 7). Still stimuli and spiky stimuli seem to have been slightly easier to learn in early generations, an early trend that leveled off while error rate continued to reduce in later generations for other stimulus types. Eventually all conditions

seemed to converge on a relatively comparable error rate when averaged across generations. Thus, in summary, the emergence of iconicity observed in shape and length measures is accompanied by an increase in learnability.

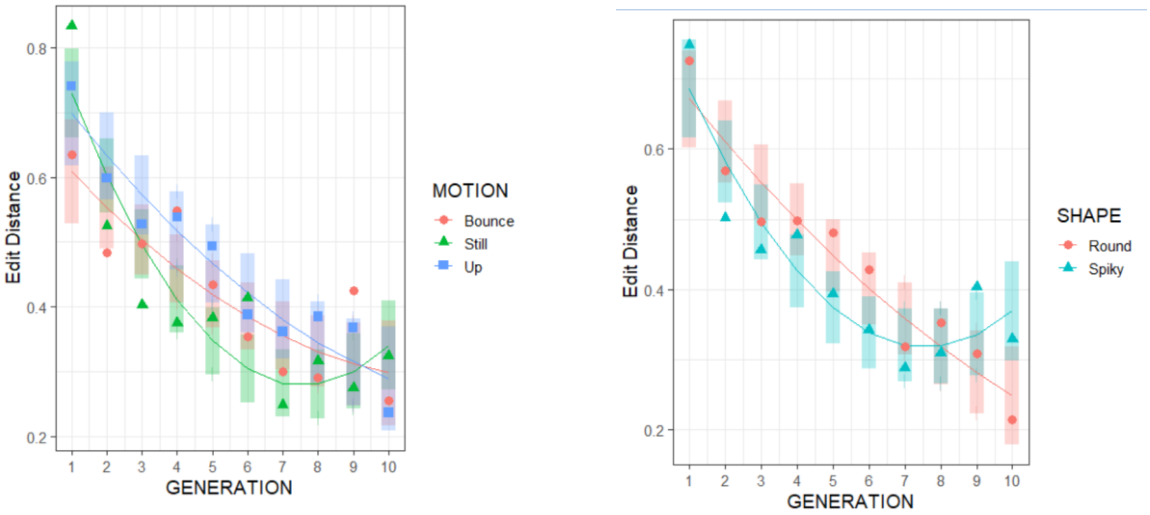

Figure 7. Error (normalized edit distance) as a function of stimulus properties and generation in Experiment 2a. Solid lines indicate model fit (linear and quadratic generation) with shaded regions indicating one standard error of model estimate at each value of generation. Solid filled points indicate observed average. Left: motion × generation; right: shape × generation

Table 3. Parameter estimates for edit distance, Experiment 2a. Reference levels: Generation = 5; Shape = mean(round, spiky); Motion = still (no motion).

	Experiment 2a Edit distance					
	Estimate	Std. Error	df	t	Pr(>|t|)	
(Intercept)	0.348	0.056	8.89	6.199	0.000	***
Generation(linear)	-0.054	0.010	12.55	-5.134	0.000	***
Generation(quadratic)	0.010	0.003	53.01	3.034	0.004	**
Shape	-0.075	0.030	897.00	-2.505	0.012	*
Motion(Bounce)	0.072	0.042	51.83	1.698	0.095	
Motion(Single)	0.119	0.043	20.30	2.781	0.011	*
Generation(linear) × Shape	0.005	0.007	897.00	0.646	0.518	
Generation(quadratic) × Shape	0.007	0.003	897.00	2.490	0.013	*
Generation(linear) × Motion(Bounce)	0.017	0.011	17.82	1.484	0.155	
Generation(quadratic) × Motion(Bounce)	-0.008	0.004	56.72	-2.048	0.045	*
Generation(linear) × Motion(Single)	0.006	0.010	25.07	0.567	0.576	
Generation(quadratic) × Motion(Single)	-0.008	0.004	54.11	-2.193	0.033	*

* p < .05 ** p < .01 *** p < 001

Exploratory Analyses

While the analyses reported in prior sections are specifically motivated by our hypotheses about shape and motion iconicity specifically, our experimental design also permits more exploratory investigation into their interrelationship, in case more complex relationships between these variables undermine the conclusions we can draw from the patterns reported above. Here,

we tested for the presence of three-way interactions between generation, shape and motion for all dependent variables in a fully factorial design, carrying out likelihood ratio tests to assess whether additional complexity offered significant improvement over the models described above. If so, we further tested whether a shape × motion × generation was warranted, by comparing the fully factorial model to one with only two-way interactions. For Experiment 1a LetterScore, the more complex model was not warranted ($\chi^2(12) = 10.724$, $p = 0.553$); three-way interactions were also not warranted in the model of edit distance ($\chi^2(6) = 4.775$, $p = 0.573$).

Generation also interacts with motion in Experiment 1a WordScore. For Experiment 1a WordScore, the interaction between shape and generation were still observed; the only additional significant predictor was an interaction between linear generation and motion that did not involve shape (illustrated in Supplementary Materials Figure S4; see Supplementary Materials Table S4 for model selection and summary statistics). Bouncing stimuli exhibited a small tendency to acquire spikier names over the generations, possibly linked to the spiky trajectories of their motion, whereas stimuli that were still or with a single upward movement acquired rounder names. These effects could also be due to an overlap between shape and motion sound symbolism. Previous work has shown that some phonemes associated with roundness/spikiness (i.e., /u/ and /i/) are also associated with slowness/quickness (Cuskley, 2013). Other work has shown that nonwords judged as being round/spiky are also judged as being still/moving (Tzeng et al., 2016).

Generation and shape interact with motion in Experiment 2b WordScore. For Experiment 2b WordScore, the observed interaction between generation and shape remained significant, but was qualified by an unexpected three-way interaction with motion (see Supplementary Materials Figure S5 and Table S5). To explore this interaction, we fit separate models for the three different motion conditions. For still and stimuli with a single upward movement, there were reliable interactions between shape and generation (still: linear generation × shape $\beta = 2.938$, SE = 1.429, $t = 2.056$, $p = 0.042$; still: quadratic generation × shape β = -

2.622, SE = 1.440, t = -1.821, p = 0.071; upward: linear generation × shape β = 2.651, SE = 1.283, t = 2.066, p = 0.040; upward: quadratic generation × shape β = -3.720, SE = 1.292, t = -2.881, p = 0.004), but for bouncing stimuli these interactions were not observed (both $|t|<1$). When we removed bouncing stimuli and fit an additional model including the three-way interaction, this interaction was no longer significant. Overall, for still and single upward movement stimuli, the divergence slowed over the generations, but for bouncing stimuli this effect of shape iconicity was not observed. This might imply that the difference between names for round and spiky stimuli was smaller when stimuli had highly salient movement which may have made shape less salient. With this exception, however, the overall pattern of divergence between spiky and round shapes over generations was maintained across both experiments.

Generation and shape interact in Experiment 2a Number of Letters. For Experiment 2a Number of Letters, the interaction between motion and generation remained significant, but there was an additional unexpected main effect of shape, qualified by an interaction with linear generation: length of round shapes did not change reliably over generations while spiky shapes developed longer names over time (See Supplementary Materials Figure S6 and Table S6). Previous work has shown that more visually complex objects tend to be given longer labels (Lewis & Frank, 2016; see also Hofer & Levy, 2019), and this phenomenon may explain the effect we observed here.

Generation and motion also interact with shape in Experiment 2b Number of Syllables. Finally, for Experiment 2b Number of Syllables, the observed interaction between generation and motion was qualified by an unexpected three-way interaction with shape (see Supplementary Materials Figure S7 and Table S7). Again, we explored the three-way interaction by fitting separate models for the three different motion conditions. For still stimuli, there was simply a linear effect of generation and no interaction with shape (linear generation β = -3.593, SE = 1.032, t = -3.480, $p < 0.001$): names for still objects became shorter over generations. For bouncing stimuli, the only significant term was a linear effect of generation in the opposite

direction (linear generation β = 5.021, SE = 1.131, t = 4.440, p < 0.001; trends for nonlinearity and interaction with shape did not reach significance): names for objects with repeated movement became longer over generations. Finally, for stimuli with a single upward movement, there was evidence for shape by generation interaction, including a quadratic term (linear generation × shape β = 3.688, SE = 1.176, t = 3.136, p = 0.002; quadratic generation × shape β = -2.600, SE = 1.174, t = -2.215, p = 0.028). As Figure S8 illustrates, spiky shapes with a single upward movement exhibited a steep increase in length in early generations, and leveled off later, while for round shapes this increase over generations was much less pronounced. This complex pattern of results may be due to slight imbalances in the initial vocabularies: round/single movement stimuli happened to have longer names in generation 0, while spiky/single movement stimuli happened to have shorter names. But despite these minor differences between conditions, we see a strong iconic relationship overall: names for motionless entities become shorter, and names for moving entities become longer over generations.

Discussion

This experiment demonstrates that iconicity for both shape and motion emerge in a model of the cultural evolution of language and is maintained through generations of speakers. Overall we found that names for rounded and spiky shapes diverged in measures of shape iconicity over generations, and this difference was maintained in later generations. And, at the same time, names for moving shapes became longer, and names for motionless shapes shorter. We also observed several novel interactions between shape and motion iconicity but which did not undermine this broad conclusion. For instance, in Experiment 2b we only observed the emergence of shape iconicity for still and one-bounce targets. This demonstrates the complexities of iconicity in existing language, with any given referent potentially affording multiple iconic mappings. An important topic for the literature going forward will be to address how languages develop and maintain iconic properties of some but not other semantic dimensions. In this experiment we observed that the affordance for motion iconicity may have overridden that of

shape iconicity. It may be that various factors serve to make some dimensions more salient than others, and that these will be more likely to elicit iconic mappings. This is particularly relevant for spoken iconicity, which may be limited in its dimensionality (Little, Eryılmaz, & de Boer, 2017), and its ability to support multiple iconic mappings (cf. Perlman et al., 2015; Perlman & Lupyan, 2018). It may also be that length is easier to iconically convey iconically than shape. That is, the link between letter length and event length may be more direct than that between phonology and shape, which requires a greater amount of mediation (via evoked associations).

General Discussion

We asked whether multiple kinds of iconicity (shape and motion) emerge when speakers need to create a novel label for a specific referent (or when they cannot remember the label and they need to produce it nonetheless) and whether they are maintained during language transmission. Experiment 1 demonstrated that iconicity emerges when participants generate labels on a single trial. Experiment 2 then used iterated learning (Kirby, et al., 2008) to show that these forms of iconicity spontaneously emerge and are maintained in a model of cultural evolution. On average, rounded stimuli developed rounder-sounding labels than spiky stimuli (e.g. Experiment 2b: “bu” vs. “tito”) and events of a longer duration developed longer labels (e.g. Experiment 2b: “kimiloka-a” vs. “bler’). In general, we found the same patterns in written and spoken modalities. The exceptions were that shape iconicity emerged quicker in Experiment 2b (speech), and the motion effect was marginal in Experiment 1b (again speech).

While shape iconicity has received substantial attention before, our results with respect to motion iconicity are the first experimental demonstration of the association between word length and event duration. This is consistent with the observations that word length is used to iconically convey event duration in languages (e.g., Perniss et al., 2010). It is important to note that differences in duration were confounded with other differences in our stimuli. For example, there were a greater number of movements in the longer duration condition; this might also serve to

make these events more complex. Thus, the relationship between length and duration could also involve an association between length and complexity (see Lewis & Frank, 2016). Future research should untangle these possibilities.

Unlike previous studies which used non-linguistic vocalizations (Perlman et al., 2015) the present results demonstrated that motion iconicity will emerge when participants must conform to an established phonology. More generally, these results are important because they are a demonstration of prosodic iconicity that can be coded into the lexicon. While other instances of prosodic iconicity emerge during language use (e.g., pitch; Perlman et al., 2015) the length of a word is more permanent (i.e., it rarely varies with language use, and is encoded in the lexicon), suggesting another way in which iconicity can affect the lexicon.

Emergence and Maintenance of Iconicity

Experiment 1 demonstrates, for the first time, shape and motion sound symbolism under completely unconstrained conditions, on a single trial. For shape, all previous demonstrations of the effect have used contrasting pairs of shapes (either in a single trial or over the course of multiple trials) and/or had participants choose from nonwords that also highlighted the relevant phonology (again, either in a single trial or over the course of the experiment). While these prior experiments certainly demonstrate that a bias exists, they leave open the question of the extent to which that bias would affect language outside of such a constrained environment. Experiment 1 clearly demonstrates the generalizability of shape sound symbolism to the simplest and most unconstrained task condition: naming a novel object.

Experiment 2 suggests that iconicity is introduced *and then maintained*. The iterated learning paradigm is an idealized model of language evolution that examines whether biases can lead to language change. The results from Experiment 2 indicate that biases toward generating iconic labels operated in the study, providing a plausible model for how iconicity might emerge

in real language. Future research might examine what advantage conferred by iconicity (e.g., greater learnability, facilitated processing), contributes to this bias. Importantly, the amount of iconicity in the lexica did not increase indefinitely. This is in line with the view, spelled out in the introduction, that the pressure towards iconicity is one of several acting on communicative systems, and will eventually be balanced by other pressures (e.g., towards discriminability and thus arbitrariness; Perniss & Vigliocco, 2014) leaving some stable amount of iconicity in all languages.

It is theoretically informative to compare the results we obtained to those in studies of iterated learning with pictorial signs (e.g., Fay et al., 2010; Garrod et al., 2010). In these studies participants draw simple pictures to communicate meaning rather than generating nonwords. A common pattern is that drawings are initially iconic, before becoming symbolic (i.e., arbitrary) in later rounds. This contrasts with our results, in which nonwords steadily became more iconic and remained so, rather than becoming arbitrary. An important distinction may be that the initial drawings are complex and idiosyncratic, and thus more difficult to produce each round. Thus, the shift away from iconicity in studies with visual signals may represent a shift towards a more conventional and easily reproducible signal. In our results, participants were somewhat limited to conventional and easily reproducible signals from the start, by being restricted to existing phonology. This may be the reason that we don't see the same early decline in iconicity. We can also speculate that a key difference is that pictorial signs can easily have a unimodal relationship with referents (e.g., a drawing looking like its referent), as opposed to nonwords which must often have a crossmodal relationship (e.g., the sound of the word mapping onto a the visual appearance of a shape). This latter kind of relationship will involve *diagrammatic iconicity*; that is, a relationship between signs being analogous to a relationship between referents (e.g., *bouba* is rounder than *kiki*, as a round shape is rounder than a spiky one; Perice, 1974). It is sensible to expect diagrammatic iconicity to take longer to develop (as it is less obvious), and to also be

better able to coexist with arbitrariness (as it is never wholly iconic and will involve arbitrary elements).

It is important to note that Experiment 2 suffered from the very issue that we discussed in the Introduction: participants were presented with multiple instances of stimuli varying on a limited number of dimensions, which may have highlighted certain contrasts. While this is certainly true, Experiment 1 served to demonstrate that iconic biases exist beyond contrasting stimuli. In addition, the stimuli in Experiment 2 varied along three different dimensions: shape, movement, and colour, making each contrast less salient. Further, the language stimuli in Experiment 2 were sound symbolically neutral and did not highlight any phonological contrast. Nevertheless, it would be worthwhile for future research to examine whether iconic biases can affect cultural transmission with visual stimuli that do not clearly highlight the relevant contrasts.

It is notable that our participants in Experiment 2 were all English speakers, and that English has not been noted for containing motion iconicity in its orthography nor phonology (for English shape iconicity see Sidhu et al., 2020). Despite this, our participants demonstrated motion iconicity. This suggests that the emergence and maintenance of motion iconicity is not dependent on experience with a language that contains such a mapping. Of course, our task involved a small number of items; previous work has shown that pressure towards arbitrariness can emerge with larger vocabularies (e.g., Gasser, 2004).

The emergence of iconicity could have implications for language origins. Scholars have speculated that the origin of language involved an iconic proto-language that mimicked absent objects and events to establish *displacement* – the power to talk about things beyond the here and now (Bickerton, 2009; Kendon, 1991). Many scholars have argued that language must originate with gesture (i.e., *gesture first theoires*; Arbib, Liebal, & Pika, 2008; Armstrong & Wilcox, 2007; Corballis, 2009; Sterelny, 2012; Tomasello, 2008). This is because, in contrast to speech, gestures have vast iconic potential, readily representing object shape, manner of motion, and

spatial relations. Conversely, iconic speech seems limited to acoustic imagery (e.g. animal sounds). For example, Sterelny (2012) claimed that "only through vocal imitation does language afford the option of a natural correspondence between sound and object, and few referents make a unique sound that humans can easily mimic" (p. 2144). Therefore, the use of vocalization might have come later (to "free" the hands) once symbolic communication was already established. However, here we have added to recent work showing that speech has more potential for iconicity than is often credited (e.g., Perlman et al., 2015; Perlman & Lupyan, 2018). Equally importantly, we have shown that iconic segmental phonology is spontaneously used to label new objects and events. Therefore our findings question the necessity of assuming that gestural systems came first in any iconic language evolution scenario.

Gesture-first theories have problems explaining how and why gestural language would subsequently be replaced by spoken language (Corballis, 2009; Kendon, 2004). Finding that vocal iconicity in phonemes/orthography (as here), or in prosody (Perlman, et al., 2015; Perlman & Lupyan, 2018) can support the creation of communicative systems, raises the possibility that both channels might have coevolved. If gesture and speech did coevolve, and assuming that arbitrariness and iconicity represent adaptations to different constraints, some form of "division of labour" between speech and gesture is plausibly optimal (Perry, Perlman, Winter, Massaro, & Lupyan, 2017; Roberts, Lewandowski, & Galantucci, 2015).

Conclusions

This study asked whether iconicity will affect the creation of new labels, and whether it will be maintained in the course of the cultural evolution of language. For the dimensions of shape and motion, we found that the answer is yes. Tentatively, these results have implications for our understanding of language origins, providing proof-of-concept evidence that iconicity (at

the segmental and supra-segmental level) can help establish spoken symbolic communicative systems, challenging the central assumption of gesture-first theories of language evolution. Thus, rather than being ornamental, iconicity may be linked to the key pressure for our communicative system to be referential, one of the central constraints operating phylogenetically and ontogenetically on language. This pressure would cause iconicity to accumulate if it is sparse, until the point where further gains from additional iconicity would be outweighed by countervailing pressures like discriminability. This theoretical view places iconicity on a footing with other language universals which cultural evolution plays a part in maintaining (Kirby, et al., 2004) – giving us an explanation of its ubiquity, and meaning that it is something we should expect to see maintained in all natural languages.

Ethics: Experiments were approved by University College London's ethics committee. Informed consent was obtained from participants (or guardians where children participated).

Data Accessibility: Data and analysis scripts (Vigliocco & Vinson, 2020) are located at: https://osf.io/ysx8t/?view_only=26b67f0426a34bb48994a2a1cbb21117

Author Contributions: GV, ALZ, and MJ developed concept. MJ, DV, ALZ, JS, and GV developed design. MJ and ALZ conducted norming. MJ programmed, produced stimuli, and collected data. MJ and DV conducted statistical analyses. MJ, GV, DS and JS drafted the manuscript; all authors provided revisions.

Competing Interests: We have no competing interests.

Acknowledgments Supported by ESRC grants RES-062-23-2012 to GV & RES-620-28-6002 to UCL DCAL; and Spanish Ministry of Economy and Competitivity grant PSI2012-32464 to JS & GV.

References

Alpher, B. (2001). Ideophones in interaction with intonation and the expression of new information in some indigenous languages of Australia. . In F.K.E. Voeltz & C. Kilian-Hatz (Eds.), *Ideophones* (pp. 9–24). Amsterdam: John Benjamins.

Arbib, MA, Liebal, K., & Pika, S. (2008). Primate vocalization, gesture, and the evolution of human language. *Curr Anthropol, 49*(6), 1053–1076.

Armstrong, D.F, & Wilcox, S. (2007). *The gestural origin of language* Oxford: Oxford Univ Press.

Asano, M., Imai, M., Kita, S., Kitajo, K., Okada, H., & Thierry, G. (2015). Sound symbolism scaffolds language development in preverbal infants. *Cortex, 63*(0), 196–205.

Atoda, T., & Hoshino, K. (1995). *Giongo Gitaigo Tsukaikata Jiten [Usage Dictionary of Sound/Manner Mimetics]* Tokyo: Sotakusha.

Barr, D.J., Levy, R., Scheepters, C., & Tily, H.J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J Mem Lang, 68*, 255–278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7. Retrieved from http://CRAN.R-project.org/package=lme4

Bickerton, D. (2009). *Adam's Tongue: How Humans Made Language, How Language Made Humans*. New York, NY: Hill & Wang.

Bremner, A.J., Caparos, S., Davidoff, J., de Fockert, J., Linell, K.J., & Spence, C. (2012). “Bouba” and “Kiki” in Namibia? A remote culture make similar shape-sound matches, but different shape-taste matches to Westerners. *Dev Sci, 9*(3), 316–322.

Cappe, C., & Rouiller, E.M.Barone, P. (2009). Multisensory anatomical pathways. *Hear Res, 258*, 28–36.

Caselli, N. K., & Pyers, J. E. (2017). The road to language learning is not entirely iconic: Iconicity, neighborhood density, and frequency facilitate acquisition of sign language. *Psychological Science*, *28*, 979-987.

Childs, G.T. (1994). African Ideophones. In L. KHinton, J. Nichols & J.J. Ohala (Eds.), *Sound Symbolism* (pp. 178–206). Cambridge: Cambridge Univ Press.

Christiansen, M.H., & Chater, N. (2008). Language as shaped by the brain. *Behav Brain Sci, 14*(9), 489–558.

Corballis, M.C. (2009). The evolution of language. *Ann NY Acad Sci, 1156*, 19–43.

Cuskley, C. (2013). Mappings between linguistic sound and motion. *Public Journal of Semiotics, 5*(1), 39-62.

De Saussure, F. (1983/1916). *Course in General Linguistics*. La Salle, IL: Open Court.

Diffloth, G. (1972). The notes on expressive meaning. In P.M. Peranteau, J.N. Levi & G.C. Phares (Eds.), *Papers from the Eighth Regional Meeting of Chicago Linguistic Society* (pp. 440–447). Chicago: Chicago Linguistic Society.

Dingemanse, M. (2011). Ezra Pound among the Mawu: Ideophones and iconicity in Siwu. In P. Michelucci, O. Fischer & C. Ljungberg (Eds.), *Semblance and Signification* (pp. 39–54). Amsterdam: John Benjamins.

Dingemanse, M., Schuerman, W., Reinisch, E., Tufvesson, S., & Mitterer, H. (2016). What sound symbolism can and cannot do: Testing the iconicity of ideophones from five languages. *Language, 92*(2), e117-e133.

Gasser, M. (2004). The origins of arbitrariness in language. *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, 4-7. Retrieved from http://www.cs.indiana.edu/l/www/pub/gasser/cogsci04.pdf

Hockett, C.F. (1960). The Origin of Speech. *Sci Am, 203*, 88–111.

Hofer, M., & Levy, R. P. (2019). Iconicity and Structure in the Emergence of Combinatoriality. https://doi.org/10.31234/osf.io/vsjkt

Ibarretxe-Antuñano, I. (2006). *Sound Symbolism and Motion in Basque*. Munich: Lincom Europa.

Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition, 109*, 54-65.

Kanero, J., Imai, M., Okuda, J., Okada, H., & Matsuda, T. (2014). How sound symbolism is processed in the brain: A study on Japanese mimetic words. *PLoS ONE*, *9*, 1-8.

Kantartzis, K., Imai, M., & Kita, S. (2011). Japanese sound-symbolism facilitates word learning in English-speaking children. *Cognitive Science, 35*, 575-586.

Kendon, A. (1991). Some Considerations for a Theory of Language Origins. *Man*, *26*(2), 199–221.

Kendon, A. (2004). *Gesture: visible actions as utterance* Cambridge: Cambridge Univ Press.

Kirby, Simon, Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proc Natl Acad Sci USA 104*(12), 5241–5245.

Kirby, Simon, Smith, Kenny, & Brighton, H. (2004). From UG to universals: linguistic adaptation through iterated learning. *Studies in Language*, *28*(3), 587–607.

Köhler, W. (1929). *Gestalt Psychology*. New York: Liveright.

Kovic, V., Plunkett, K., & Westermann, G. (2010). The shape of words in the brain. *Cognition, 114*, 19-28.

Kuznetsova A., Brockhoff, P. B., & Christensen, R. H. B. (2016). lmerTest: Tests in Linear Mized Effects Models. (Version 2.0-33) [R Package]. Retrieved from https://CRAN.R-project.org/package=lmerTest

Laing, C.E. (2014). A phonological analysis of onomatopoeia in early word production. *First Lang, 34*(5), 387–405.

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

Locke, J. (1690). *Essay Concerning Human Understanding*. London: Thomas Bassett.

Lockwood, G., & Dingemanse, M. (2015). Iconicity in the lab: a review of behavioural, developmental, and neuroimaging research into sound-symbolism. *Front Psychol, 6*, 1246.

Lockwood, G., Dingemanse, M., & Hagoort, P. (2016). Sound-symbolism boosts novel word learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*, 1274.

Lockwood, G., Hagoort, P., & Dingemanse, M. (2016). How Iconicity Helps People Learn New Words : Neural Correlates and Individual Differences in Sound-Symbolic Bootstrapping, *Collabra, 2*, 1–15.

Maurer, D., Pathman, T., & Mondloch, C.J. (2006). The shape of boubas: sound-shape correspondences in toddlers and adults. *Dev Sci, 9*(3), 316–322.

McCormick, K., Kim, J. Y., List, S., & Nygaard, L. C. (2015). Sound to meaning mappings in the bouba–kiki effect. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1565–1570). Austin, TX: Cognitive Science Society.

Meteyard, L., Stoppard, E., Snudden, D., Cappa, S.F., & Vigliocco, G. (2015). When semantics aids phonology: A processing advantage for iconic word forms in aphasia. *Neuropsychologia, 76*, 264–275.

Mikone, E. (2001). Ideophones in the Balto-Finnic Languages. In F.K.E. Voeltz & C. Kilian-Hatz (Eds.), *Ideophones* (pp. 223–233). Amsterdam: John Benjamins.

Monaghan, P. , Christiansen, M.H., & Fitneva, S.A. (2011). The arbitrariness of the sign: Learning advantages from the structure of the vocabulary. *J Exp Psychol Gen, 140*, 325–347.

Monaghan, P., Shillcock, R.C., Christiansen, M.H., & Kirby, Simon. (2014). How Arbitrary is Language? *Philos Trans R Soc Lond B Biol Sci, 369*, 20130299.

Nielsen, A. K., & Dingemanse, M. (2020). Iconicity in Word Learning and Beyond: A Critical Review. *Language and Speech*. doi: 10.1177/0023830920914339

Nielsen, A., & Rendall, D. (2012). The source and magnitude of sound-symbolic biases in processing artificial word material and their implications for language learning and transmission. *Language and Cognition, 4*, 115-125.

Nielsen, A. K., & Rendall, D. (2013). Parsing the role of consonants versus vowels in the classic Takete-Maluma phenomenon. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 67*, 153-163.

Nuckolls, J. (1996). *Sounds Like Life*. New York: Oxford Univ Press.

Ohala, JJ. (1994). The frequency code underlies the sound-symbolic use of voice pitch. In L. Hinton, J. Nichols & J.J. Ohala (Eds.), *Sound Symbolism* (pp. 325–347). Cambridge: Cambridge Univ Press.

Ozturk, O, Krehm, M., & Vouloumanos, A. (2013). Sound symbolism in infancy: Evidence for sound-shape cross-modal correspondences in 4-month-olds. *J Exp Child Psychol, 114*(2), 173–186.

Pejovic, J., & Molnar, M. (2017). The development of spontaneous sound-shape matching in monolingual and bilingual infants during the first year. *Developmental Psychology, 53*, 581-586.

Perlman, M., Dale, R., & Lupyan, G. (2015). Iconicity can ground the creation of vocal symbols. *Royal Soc Open Sci 2*, 150152.

Perlman, M., & Lupyan, G. (2018). People can create iconic vocalizations to communicate various meanings to naïve listeners. *Scientific Reports, 8*(1), 2634.

Perniss, P., & Vigliocco, G. (2014). The bridge of iconicity: From a world of experience to the experience of language. *Philos Trans R Soc Lond B Biol Sci 369*, 20130300.

Perry, L K, Perlman, M, Winter, B, Massaro, D W, & Lupyan, G. (2017). Iconicity in the speech of children and adults. *Developmental Science*. doi: 10.1111/desc.12572

Perry, L.K., Perlman, M., & Lupyan, G. (2015). Iconicity in English and Spanish and its relation to lexical category and age of acquisition. *PLoS ONE, 109*, e0137147.

Pietrandrea, P. (2002). Iconicity and arbitrariness in Italian Sign Language. *Sign Language Studies, 2*(3), 296-321.

R_Core_Team. (2014). *R : A language and environment for statistical computing*: R Foundation for Statistical Computing, Vienna, Austria.

Ramachandran, V.S., & Hubbard, E.M. (2001). Synesthesia: a window into perception, thought and language. *J Conscious Stud, 8*(12), 3–34.

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

Scott-Phillips, T., & Kirby, Simon. (2010). Language evolution in the laboratory. *Trends Cogn Sci, 14*(9), 411–417.

Sereno, M. (2014). Origin of symbol-using systems: speech, but not sign, without the semantic urge. *Philosophical Transactions of the Royal Society B, 369*, 20130303.

Shintel, H., Nusbaum, H.C., & Okrent, A. (2006). Analog acoustic expression in speech communication. *J Mem Lang, 55*, 167–177.

Sidhu, D. M., & Pexman, P. M. (2018). Lonely sensational icons: semantic neighbourhood density, sensory experience and iconicity. *Language, Cognition and Neuroscience, 33*, 25-31.

Sidhu, D. M., Westbury, C., Hollis, G., & Pexman, P. M. (accepted pending revisions). Sound symbolism shapes the English language: The maluma/takete effect in English nouns. *Psychonomic Bulletin & Review*.

Sterelny, K. (2012). Language, gesture, skill: The co-evolutionary foundations of language. *Philos Trans R Soc Lond B Biol Sci, 367*(1599), 2141–2151.

Tamariz, M., Roberts, S., Martinez, I., & Santiago, J. (2018). The interactive origin of iconicity. *Cognitive Science, 42,* 334-349. doi: 10.1111/cogs.12497

Tomasello, M. (2008). *Origins of Human Communication*. Cambridge, MA: MIT Press.

Tzeng, C. Y., Nygaard, L. C., & Namy, L. L. (2017). The specificity of sound symbolic correspondences in spoken language. *Cognitive Science, 41*, 2191-2220.

Verhoef, T., Roberts, S. G. & Dingemanse, M. (2015) Emergence of systematic iconicity: Transmission, interaction and analogy. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. (pp. 2481-2486) Austin,TX: Cognitive Science Society.

Vigliocco, G., & Kita, S. (2006). Language specific effects of meaning, sound and syntax: implications for models of lexical retrieval in production. *Lang Cogn Process, 21*, 790–816.

Vigliocco, G., & Vinson, D. (2020, June 29). Iconicity emerges and is maintained in spoken language. Retrieved from osf.io/ysx8t

Vinson, D. P., Cormier, K., Denmark, T., Schembri, A., & Vigliocco, G. (2008). The British Sign Language (BSL) norms for age of acquisition, familiarity, and iconicity. *Behavior Research Methods, 40*(4), 1079-1087.

Watson, R.L. (2001). A comparison of some Southeast Asian ideophones with some African ideophones. In F.K.E. Voeltz & C. Kilian-Hatz (Eds.), *Ideophones* (pp. 385–405). Amsterdam: John Benjamins.

Westbury, C. (2005). Implicit sound symbolism in lexical access: evidence from an interference task. *Brain and Language, 93*(10-19).

Supplementary Tables and Figures

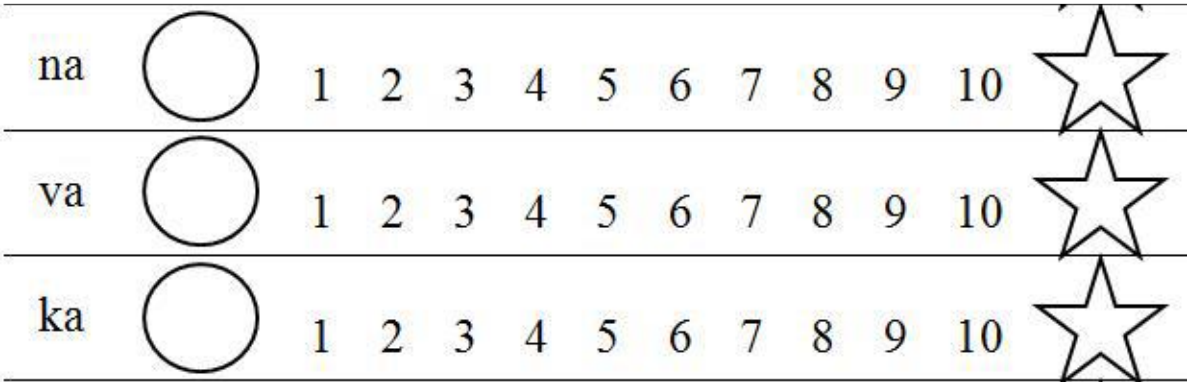

Figure S1. The rating scale used for the norming that produced the LetterScore metric.

Figure S2. The rating scale used for the norming that produced the WordScore metric (reversed for half the participants).

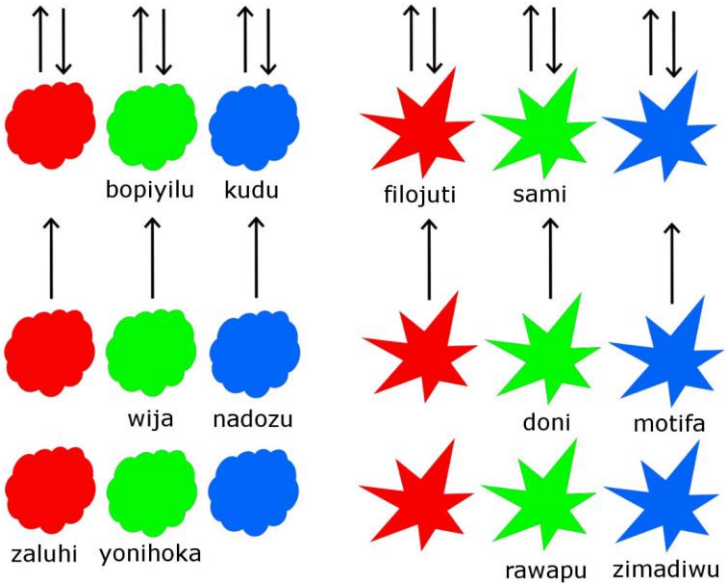

Figure S3. Complete set of items used in Experiment 2a. Arrows depict movement and were not visible to participants. Upper row: stimuli moved with a bouncing motion. Middle row: stimuli moved with a single upward motion. Lower row: stimuli did not move. Labels on some items depict Seen items, for which labels were assigned in generation 0. Items without labels were not seen by participants in generation 1.

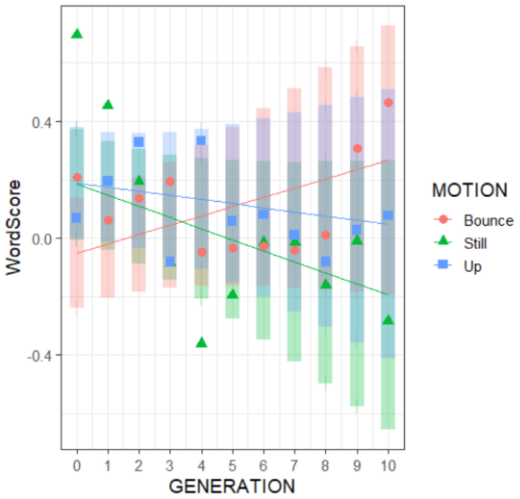

Figure S4. Wordscore as a function of stimulus motion and generation in Experiment 2a. Higher values indicate more spiky labels and lower values more rounded. Solid lines indicate model fit (linear and quadratic generation) with shaded regions indicating one standard error of model estimate at each value of generation. Solid filled points indicate observed average.

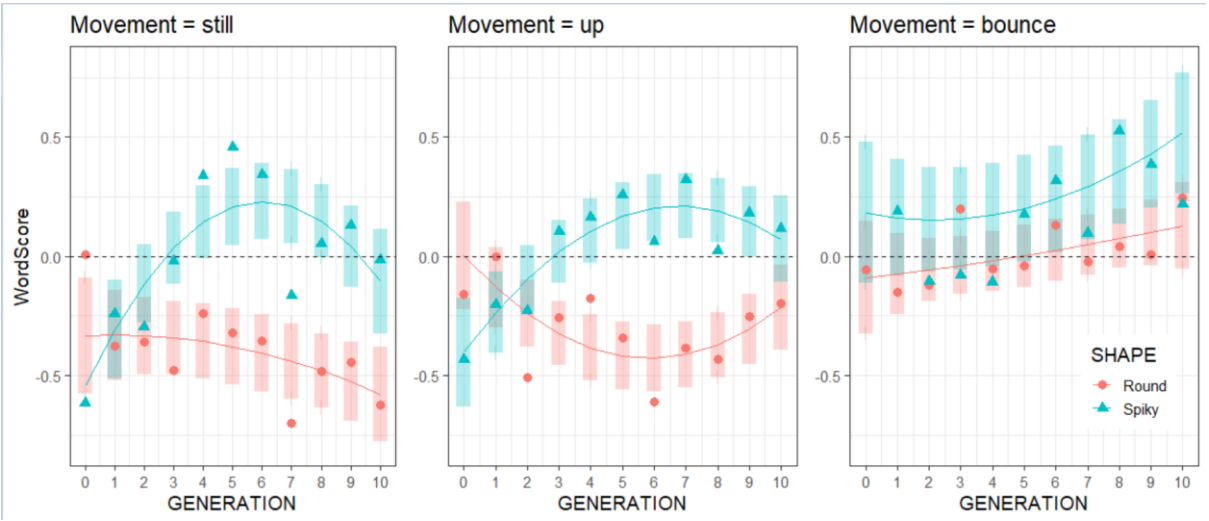

Figure S6. WordScore as a function of stimulus shape, movement and generation in Experiment 2b. Solid lines indicate model fit (linear and quadratic generation) with shaded regions indicating one standard error of model estimate at each value of generation. Solid filled points indicate observed average.

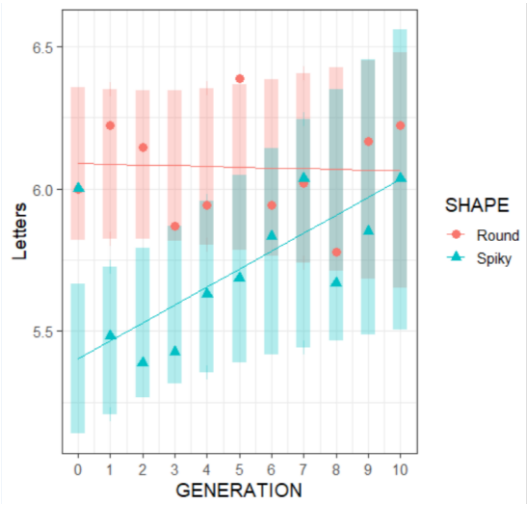

Figure S7. Number of letters as a function of stimulus shape and generation in Experiment 2a. Solid lines indicate model fit (linear and quadratic generation) with shaded regions indicating one standard error of model estimate at each value of generation. Solid filled points indicate observed average.

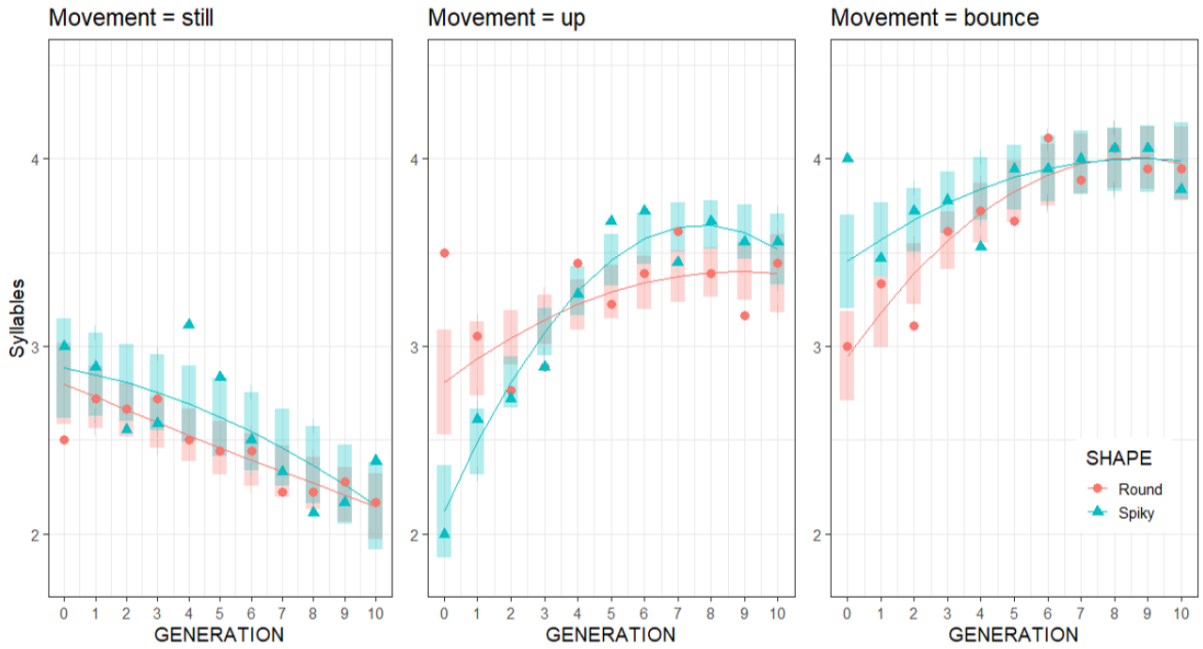

Figure S8. Number of syllables as a function of stimulus shape, movement and generation in Experiment 2b. Solid lines indicate model fit (linear and quadratic generation) with shaded regions indicating one standard error of model estimate at each value of generation. Solid filled points indicate observed average

Table S1. CV syllable scores (centered). Higher values indicate more spiky, lower values more round. See main text for further details.

CV	Score	CV	Score	CV	Score
ba	-0.72	ku	0.89	so	-1.47
be	-0.47	la	-0.83	su	-1.08
bi	0.21	le	-0.68	ta	1.00
bo	-1.58	li	0.82	te	1.67
bu	-0.90	lo	-1.75	ti	1.75
da	-0.40	lu	-1.50	to	-0.18
de	0.67	ma	-1.54	tu	0.50
di	0.96	me	-0.86	va	0.35
do	-1.40	mi	-0.15	ve	0.67
du	-1.08	mo	-2.22	vi	1.67
fa	0.07	mu	-1.86	vo	-0.54
fe	0.63	na	-0.18	vu	0.28
fi	1.17	ne	0.07	wa	0.25
fo	-0.83	ni	1.03	we	0.17
fu	-0.50	no	-0.83	wi	0.10
ha	0.10	nu	-0.43	wo	-1.58
he	-0.43	pa	-0.61	wu	-1.18
hi	0.00	pe	0.85	ya	0.50
ho	-1.61	pi	1.04	ye	0.28
hu	-0.86	po	-0.75	yi	1.60
ja	0.39	pu	0.07	yo	-1.54
je	0.57	ra	-0.33	yu	-0.85
ji	1.53	re	0.57	za	1.35
jo	-0.72	ri	0.71	ze	1.92
ju	-1.08	ro	-0.79	zi	1.78
ka	2.03	ru	-0.36	zo	0.21
ke	2.21	sa	0.00	zu	0.75
ki	2.35	se	0.25		
ko	0.53	si	0.46		

Table S2. LetterScores (centered). Higher values = more spiky. * Letters c, g, q, x were not included in syllable norming and were assigned values of zero, see main text for details.

a	0.08	k	1.60	u	-0.54
b	-0.70	l	-0.79	v	0.48
c	0*	m	-1.33	w	-0.45
d	-0.25	n	-0.07	x	0*
e	0.47	o	-1.01	y	-0.01
f	0.10	p	0.12	z	1.20
g	0*	q	0*		
h	-0.56	r	-0.04		
i	1.00	s	-0.37		
j	0.13	t	0.94		

Table S3. Transcription of the original language (generation 0) from Experiment 1b. The same shapes were used as in Experiment 1a (see Figure S3).

Shape	**Colour**	**Motion**	**Name**
Round	Blue	Still	[haɪ'maɪwaɪ]
Round	Blue	Up	[ni'luvaɪ]
Round	Green	Bounce	[kiwəʊ'mudʒaɪ]
Round	Green	Up	[dəʊpaɪ'həʊkɑ]
Round	Red	Bounce	['wuzɑ]
Round	Red	Still	['təʊfɑ]
Spiky	Blue	Bounce	[zidʒəʊ'taɪjəʊ]
Spiky	Blue	Up	['zaɪməʊ]
Spiky	Green	Bounce	[səʊti'kaɪbəʊ]
Spiky	Green	Still	[faɪ'mɑnɑ]
Spiky	Red	Still	[ləʊjaɪ'pu]
Spiky	Red	Up	[hɑ'sɑ]

Table S4. Summary of results for exploratory analysis of Experiment 2a Wordscore. Reference levels: Generation = 5; Shape = mean(round, spiky); Motion = still (no motion). Full factorial model vs. hypothesis-driven model (see main text): $\chi^2(12) = 26.71$, $p = 0.009$; Full factorial model vs. model with only two-way interactions: $\chi^2(6) = 4.70$, $p = 0.582$. Model with only two-way interactions (including motion) was retained.

Experiment 2a WordScore (includes two-way interactions with motion)

	Estimate	Std. Error	df	t	Pr(>|t|)	
(Intercept)	-0.145	0.186	6.97	-0.781	0.461	
Generation(linear)	-0.060	0.044	7.44	-1.350	0.217	
Generation(quadratic)	0.015	0.007	33.95	2.211	0.034	*
Shape	0.331	0.136	16.80	2.445	0.026	*
Motion(Bounce)	0.100	0.090	1014.00	1.108	0.268	
Motion(Single)	0.223	0.090	1014.00	2.469	0.014	*
Generation(linear) × Shape	0.105	0.023	10.03	4.634	0.001	***
Generation(quadratic) × Shape	-0.020	0.007	176.80	-3.033	0.003	**
Generation(linear) × Motion(Bounce)	0.068	0.022	1014.00	3.121	0.002	**
Generation(quadratic) × Motion(Bounce)	0.002	0.008	1014.00	0.197	0.844	
Generation(linear) × Motion(Single)	0.034	0.022	1014.00	1.542	0.124	
Generation(quadratic) × Motion(Single)	-0.012	0.008	1014.00	-1.494	0.136	

Table S5. Summary of results for exploratory analysis of Experiment 2b Wordscore. Reference levels: Generation = 5; Shape = mean(round, spiky); Motion = still (no motion). Full factorial model vs. hypothesis-driven model (see main text): $\chi^2(12) = 65.01$, $p < 0.0001$; Full factorial model vs. model with only two-way interactions: $\chi^2(6) = 20.69$, $p = 0.0021$. Model with three-way interactions was retained.

Experiment 2b WordScore (includes three-way interaction)

	Estimate	Std. Error	df	t	Pr(>|t|)	
(Intercept)	-0.081	0.106	13.32	-0.761	0.460	
Generation(linear)	0.010	0.018	120.56	0.518	0.605	
Generation(quadratic)	-0.012	0.007	121.18	-1.807	0.073	
Shape	0.648	0.199	11.02	3.250	0.008	**
Motion(Bounce)	0.142	0.075	958.30	1.891	0.059	
Motion(Single)	-0.031	0.075	958.30	-0.417	0.677	
Generation(linear) × Shape	0.072	0.026	348.40	2.736	0.007	**
Generation(quadratic) × Shape	-0.020	0.010	348.29	-2.035	0.043	*
Generation(linear) × Motion(Bounce)	0.016	0.018	958.30	0.881	0.379	
Generation(quadratic) × Motion(Bounce)	0.016	0.007	958.30	2.457	0.014	*
Generation(linear) × Motion(Single)	0.004	0.018	958.30	0.193	0.847	
Generation(quadratic) × Motion(Single)	0.012	0.007	958.30	1.757	0.079	
Shape × Motion(Bounce)	-0.539	0.150	958.30	-3.592	0.000	***
Shape × Motion(Single)	-0.039	0.150	958.30	-0.259	0.795	
Generation(linear) × Shape × Motion(Bounce)	-0.066	0.036	958.30	-1.828	0.068	
Generation(quadratic) × Shape × Motion(Bounce)	0.027	0.013	958.30	2.048	0.041	*
Generation(linear) × Shape × Motion(Up)	-0.001	0.036	958.30	-0.036	0.972	
Generation(quadratic) × Shape × Motion(Up)	-0.007	0.013	958.30	-0.499	0.618	

Table S6. Summary of results for exploratory analysis of Experiment 2a Number of Letters. Reference levels: Generation = 5; Shape = mean(round, spiky); Motion = still (no motion). Full factorial model vs. hypothesis-driven model (see main text): $\chi^2(12) = 25.74$, $p = 0.0022$; Full factorial model vs. model with only two-way interactions: $\chi^2(6) = 4.79$ $p = 0.571$. Model with only two-way interactions (including shape) was retained.

Experiment 2a Length (includes two-way interactions with shape)

	Estimate	Std. Error	df	t	Pr(>|t|)	
(Intercept)	5.445	0.342	4.53	15.938	0.000	***
Generation(linear)	-0.095	0.056	6.20	-1.696	0.139	
Generation(quadratic)	0.015	0.015	6.83	1.042	0.333	
Shape	-0.329	0.133	974.26	-2.473	0.014	*
Motion(Bounce)	0.516	0.392	4.16	1.318	0.255	
Motion(Single)	0.709	0.485	4.32	1.464	0.212	
Generation(linear) × Shape	0.070	0.032	974.26	2.170	0.030	*
Generation(quadratic) × Shape	-0.007	0.012	974.26	-0.625	0.532	
Generation(linear) × Motion(Bounce)	0.239	0.086	6.16	2.795	0.031	*
Generation(quadratic) × Motion(Bounce)	0.003	0.027	6.32	0.105	0.920	
Generation(linear) × Motion(Single)	0.127	0.075	6.38	1.697	0.138	
Generation(quadratic) × Motion(Single)	-0.036	0.021	6.85	-1.718	0.131	

Table S7. Summary of results for exploratory analysis of Experiment 2b Number of Syllables. Reference levels: Generation = 5; Shape = mean(round, spiky); Motion = still (no motion).
Full factorial model vs. hypothesis-driven model (see main text): $\chi^2(12) = 28.59$, $p = 0.0008$; Full factorial model vs. model with only two-way interactions: $\chi^2(6) = 19.23$, $p = 0.0038$. Model with three-way interactions was retained.

Experiment 2b Length (includes 3-way interaction)

	Estimate	Std. Error	df	t	Pr(>|t|)	
(Intercept)	2.524	0.163	7.23	15.487	0.000	***
Generation(linear)	-0.072	0.023	6.65	-3.126	0.018	*
Generation(quadratic)	-0.002	0.005	50.00	-0.319	0.751	
Shape	0.164	0.091	895.74	1.801	0.072	.
Motion(Bounce)	1.335	0.232	6.90	5.746	0.001	***
Motion(Single)	0.852	0.212	7.53	4.013	0.004	**
Generation(linear) × Shape	-0.007	0.022	895.43	-0.310	0.756	
Generation(quadratic) × Shape	-0.005	0.008	895.23	-0.626	0.532	
Generation(linear) × Motion(Bounce)	0.148	0.045	6.46	3.272	0.015	*
Generation(quadratic) × Motion(Bounce)	-0.010	0.007	49.08	-1.417	0.163	
Generation(linear) × Motion(Single)	0.168	0.037	6.56	4.501	0.003	**
Generation(quadratic) × Motion(Single)	-0.016	0.008	49.15	-2.091	0.042	*
Shape × Motion(Bounce)	-0.099	0.129	896.05	-0.765	0.444	
Shape × Motion(Single)	0.010	0.129	895.49	0.076	0.940	
Generation(linear) × Shape × Motion(Bounce)	-0.041	0.031	896.25	-1.297	0.195	
Generation(quadratic) × Shape × Motion(Bounce)	0.012	0.011	895.97	1.041	0.298	
Generation(linear) × Shape × Motion(Up)	0.088	0.031	895.58	2.831	0.005	**
Generation(quadratic) × Shape × Motion(Up)	-0.013	0.011	895.23	-1.134	0.257	