The emergence of words from vocal imitations

Pierce Edmiston (pedmiston@wisc.edu)
Marcus Perlman (mperlman@gmail.com)
Gary Lupyan (lupyan@wisc.edu)

Abstract

We investigated how conventional spoken words might emerge from imitations of environmental sounds. Participants played a version of the children's game "Telephone". The first generation of participants imitated recognizable environmental sounds (e.g., glass breaking, water splashing). Subsequent generations imitated the imitations of the prior generation for a maximum of 8 generations. The results showed that the imitations became more stable and word-like, and more easily learnable as category labels. At the same time, even after 8 generations, both spoken imitations and their written transcriptions could be matched above chance to the category of environmental sound that motivated them. These results show how repeated imitation can create progressively more word-like forms that continue to retain a resemblance to the original sound that motivated them. The results speak to the possible role of human vocal imitation in explaining the origins of spoken words.

People have long pondered the origins of languages, especially the words that compose them. For example, both Plato in his *Cratylus* dialogue (Plato & Reeve, 1999) and John Locke in his *Essay Concerning Human Understanding* (Locke, 1948) examined the "naturalness" of words—whether they are somehow imitative of their meaning. Here we investigated whether new words can be formed from the repetition of non-verbal vocal imitations. Does the repetition of imitations over generations of speakers gradually give rise to novel word forms? In what ways do these words resemble the original sounds that motivated them? We report a large-scale experiment (N=1571) investigating how new words can form—gradually and without instruction—simply by repeating imitations of environmental sounds.

The importance of imitation and depiction in the origin of signs is clearly observable in the origin of words in signed languages (Goldin-Meadow, 2016; Kendon, 2014; Klima & Bellugi, 1980), but in considering the idea that vocal imitation may be key to understanding the origin of spoken words, many have argued that the human capacity for vocal imitation is far too limited to play a significant role (Arbib, 2012; Armstrong & Wilcox, 2007; Corballis, 2003; Hewes, 1973; Hockett, 1978; Tomasello, 2010). For example, Pinker & Jackendoff (2005) argued that, "most humans lack the ability… to convincingly reproduce environmental sounds… Thus 'capacity for vocal imitation' in humans might be better described as a capacity to learn to produce speech" (p. 209). Consequently, it is still widely assumed that vocal imitation—or more broadly, the use of any sort of resemblance between form and meaning—cannot be important to understanding the origin of spoken words.

Although most words of contemporary spoken languages are not clearly imitative in origin, there has been a growing recognition of the importance of imitative words in spoken languages (Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015; Perniss, Thompson, & Vigliocco, 2010) and the frequent use of vocal imitation and depiction in spoken discourse (Clark & Gerrig, 1990; Lewis, 2009). This has led some to argue for the importance of imitation for understanding the origin of spoken words (e.g., Brown, Black, & Horowitz, 1955; Dingemanse, 2014; Donald, 2016; Imai & Kita, 2014; Perlman, Dale, & Lupyan, 2015). In addition, experiments show that counter to previous assumptions, people are highly effective at using vocal imitations in reference—in some cases, even more effective than with conventional words (Lemaitre & Rocchesso, 2014). Recent work has also shown that people are able to create novel imitative vocalizations for more abstract meanings (e.g. 'slow', 'rough', 'good', 'many') that are understandable to naïve listeners (Perlman et al., 2015). The effectiveness of these imitations arises not because people can mimic environmental sounds with high fidelity, but because they are able to produce imitations that capture the salient features of sounds in ways that are understandable to listeners (Lemaitre, Houix, Voisin, Misdariis, & Susini, 2016). Similarly, the features of onomatopoeic words might highlight distinctive aspects of the sounds they represent. For example, the initial voiced, plosive /b/ in "boom" represents an abrupt, loud onset, the back vowel /u/ a low pitch, and the nasalized /m/ a slow, muffled decay (Rhodes, 1994).

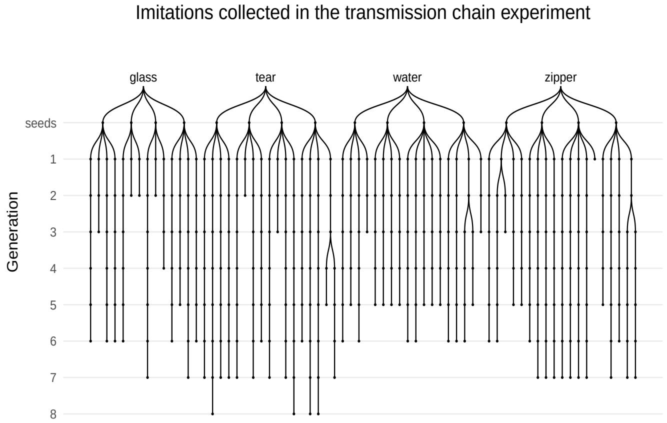


Figure 1: The design of the transmission chain experiment. Seed sounds (16) were sampled from four categories of environmental sounds: glass, tear, water, zipper. Participants imitated each seed sound, and then the next generation of participants imitated the imitations and so on for up to 8 generations. Chains are unbalanced due to random assignment and the exclusion of some low quality recordings.

Thus, converging evidence suggests that people can use vocal imitation as an effective means of communication. But can vocal imitations give rise to words that can be integrated into the vocabulary of a language? And if so, what is required for this to happen? What happens to a vocal imitation in the course of it being turned into a word? To answer these questions, we recruited participants to play an online version of the children's game of "Telephone". In the children's game, a spoken message is whispered from one person to the next. In our version, the original message or "seed sound" was a recording of an environmental sound. The initial group (first generation) of participants imitated these seed sounds, the next generation imitated the previous imitators, and so on for up to 8 generations (Fig. 1).

In subsequent experiments, we systematically answered the following questions about the form of the vocalizations and their potential to function as words. First, does iterated imitation drive the vocalizations to stabilize in form and become more word-like? Second, do the imitations become more suitable as labels for the category of sounds that motivated them? For example, does the imitation of a particular water-splashing sound become, over time, a better label for the more general category of water-splashing sounds? Third, do the imitations retain a resemblance to the original environmental sounds that inspired them? If so, it should be possible for naïve participants to match the emergent words back to the seed sounds that were originally imitated.

Results

We begin with a summary of our main results: (1) Imitations of environmental sounds became more stable over the course of being repeatedly imitated as revealed by increasing acoustic similarity along individual transmission chains. In addition, later generations of imitations had higher levels of agreement when transcribed into English orthography further suggesting an increase in stability and word-likeness. (2) When transcriptions of first and last generation imitations were learned as novel labels for categories of environmental sounds, last generation transcriptions were learned faster and generalized to new category members more easily than transcriptions of first generation imitations, suggesting that repeating imitations caused the forms to become more suitable as category labels. (3) Even as the imitations became more word-like, they also retained a resemblance to the original category of environmental sound that motivated them, as measured by the ability of naïve listeners to match both the auditory imitations and their written transcriptions to the correct category of environmental sounds even after 8 generations of repetition.

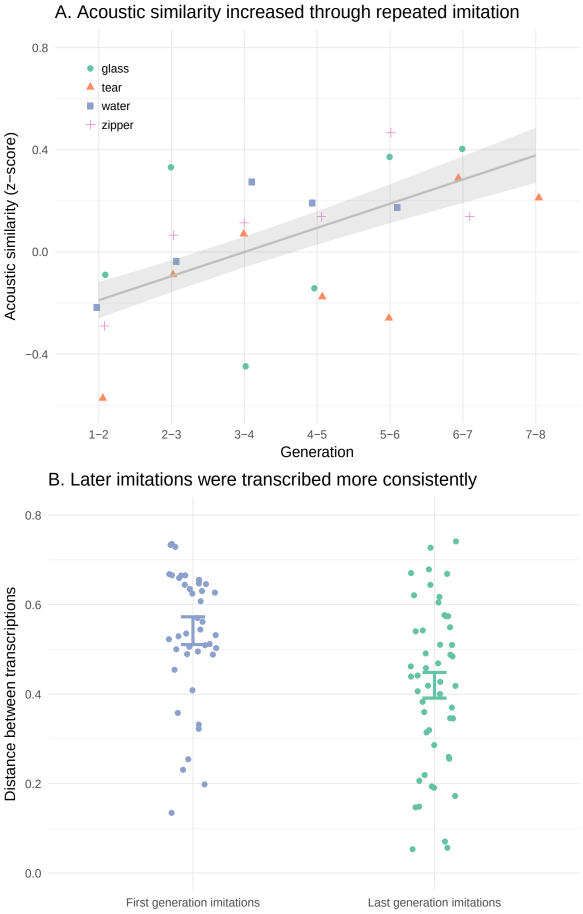


Figure 2: Stabilization of imitations through repetition. A. Change in perception of acoustic similarity over generations of repetition. Means and hierarchical linear model predictions with ±1 SE are shown. Acoustic similarity increased over generations, indicating that repetition made the vocalizations easier to imitate with high fidelity. B. Transcription similarity for first and last generation imitations. Mean orthographic distance between the most frequent transcription and all other transcriptions of a given imitation are shown, with error bars as ±1 SE of the hierarchical linear model predictions. Transcriptions of later generation imitations were more similar to one another than transcriptions of first generation imitations.

Iterated imitations became more stable and word-like

The final set of vocal imitations included 365 imitations along 105 contiguous transmission chains from 94 participants (Fig. 1; see Methods). Research assistants (N=5) rated the acoustic similarity of pairs of imitations while blind to all conditions and hypotheses (see Methods). We also conducted automated analyses of acoustic similarity using Mel Frequency Cepstral Coefficients (MFCCs) as a measure of acoustic distance. These results are reported in the Supporting Information (Fig. 9). Acoustic similarity ratings were fit with a hierarchical linear model[1] predicting similarity from generation with random effects for rater and for category. Imitations from later generations were rated as sounding more similar to one another than imitations from earlier generations, $b = 0.09$ (SE $= 0.02$), $t(4.5) = 4.42$, $p = 0.009$ (Fig. 2A). This result suggests that imitations became more stable (i.e., easier to imitate with high fidelity) with each generation.

As an additional test of stabilization, we had English-speaking participants transcribe a sample of first and last generation imitations into English orthography, and then measured whether transcription similarity

[1] Degrees of freedom and corresponding significance tests for hierarchical linear models were estimated using the Satterthwaite approximation (Kuznetsova, Bruun Brockhoff, & Haubo Bojesen Christensen, 2016).

(spelling agreement) increased over generations. We collected a total of 2163 transcriptions—approximately 20 transcriptions per sound (see Methods). Some examples of the transcriptions are presented in Table 1.

Table 1: Examples of invented words.

Category	Seed	First generation	Last generation
glass	1	tingtingting	dundunduh
glass	2	chirck	correcto
glass	3	dirrng	wayew
glass	4	boonk	baroke
tear	1	scheeept	cheecheea
tear	2	feeshefee	cheeoooo
tear	3	hhhweerrr	chhhhhhewwwe
tear	4	ccccchhhhyeaahh	shhhhh
water	1	boococucuwich	eeverlusha
water	2	chwoochwooochwooo	cheiopshpshcheiopsh
water	3	atoadelchoo	mowah
water	4	awakawush	galonggalong
zipper	1	euah	izoo
zipper	2	zoop	veeeep
zipper	3	arrgt	owww
zipper	4	bzzzzup	izzip

To measure the similarity among transcriptions of a given imitation, we calculated the orthographic distance between the most frequent transcription and all other transcriptions. The orthographic distance measure was a ratio based on longest contiguous matching subsequences between pairs of transcriptions. A hierarchical linear model predicting orthographic distance from the generation of the imitation (First generation, Last generation) with random effects for transmission chains nested within categories of environmental sounds revealed that transcriptions of last generation imitations were more similar to one another than transcriptions from first generation imitations, b = -0.12 (SE = 0.03), $t(14.5)$ = -4.15, $p < 0.001$ (Fig. 2B). The same result is reached through alternative measures of orthographic distance such as exact string matching and length of longest substring match, and when excluding imitations for which all transcriptions were unique in which case there was no most frequent transcription (Fig. 12). These results support our hypothesis that unguided repetition drives imitations to stabilize on particular words.

Iterated imitations made for better category labels

One consequence of imitations becoming more word-like is that they may make for better category labels. For example, an imitation from a later generation, by virtue of having a more word-like form, may be easier to learn as a label for the category of sounds that motivated it than an earlier imitation, which may be more closely yoked to an individual seed sound. To the extent that repeating imitations abstracts away the idiosyncrasies of a particular category member, it may also be easier to generalize to new category members. We tested these predictions using a category learning task wherein participants had to learn novel labels for categories of environmental sounds. Unbeknownst to the participants, the novel labels they learned were transcriptions generated either from first or last generation imitations. The procedure for selecting otherwise-equal transcriptions is detailed in the Supporting Information. Here we focus on the consequences of learning either first or last generation transcriptions in the category learning experiment.

At the beginning of the experiment, where participants had to learn through trial-and-error which labels were associated with which sounds, participants learning transcriptions of first or last generation imitations did not differ in overall accuracy, $p = 0.887$, or reaction time, $p = 0.616$. After this initial learning phase (i.e. after the first block of trials), accuracy performance quickly reached ceiling (Fig. 15) and did not differ

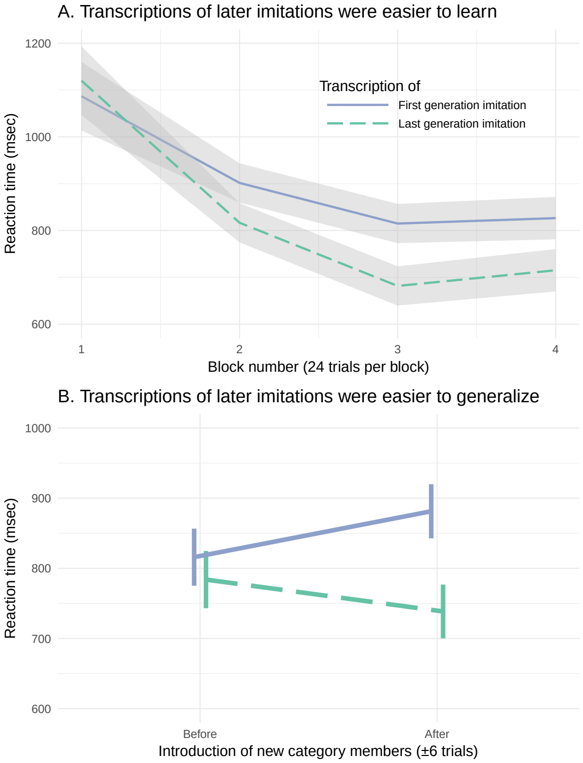


Figure 3: Learning transcriptions of imitations as category labels. A. Participants achieved faster RTs in matching transcribed labels to environmental sounds for labels transcribed from later compared to earlier generation imitations. B. There was a generalization cost for first-generation labels, but not last generation labels.

between groups $p = 0.775$. However, participants learning last generation transcriptions responded more quickly in subsequent blocks than participants learning first generation transcriptions, $b =$ -114.13 (SE = 52.06), $t(39.9) =$ -2.19, $p = 0.034$ (Fig. 3A). These faster responses suggest that, in addition to becoming more stable both in terms of acoustic and orthographic properties, repeating imitations makes them easier to learn as category labels. Given how quickly accuracy performance reached ceiling, further investigation with a more difficult category learning experiment is warranted (e.g., more than four categories and 16 exemplars).

Next, we examined whether transcriptions from last generation imitations were easier to generalize to novel sounds. To test this hypothesis, we compared RTs on trials immediately prior to the introduction of novel sounds (new category members) and the first trials after the block transition (±6 trials). The results revealed a reliable interaction between the generation of the transcribed imitation and the block transition, $b =$ -110.77 (SE = 52.84), $t(39.7) =$ -2.10, $p = 0.042$ (Fig. 3B). This result suggests that transcriptions from later generation imitations were easier to generalize to new category members.

Iterated imitations retained resemblance to original sounds

As the imitations became more word-like, were they stabilizing on arbitrary acoustic and orthographic forms, or did they maintain some resemblance to the original environmental sound that motivated them? To test this, we measured the ability of participants naïve to the design of the experiment to match imitations back to their original source relative to other seed sounds from either the same category or from different categories (Fig. 4A). All 365 imitations were tested in the three question types depicted in Fig. 4A. These questions differed in the relationship between the imitation and the four seed sounds provided as the choices in the question. Responses were fit by a hierarchical generalized linear model predicting match accuracy as different from chance (25%) based on the type of question being answered (True seed, Category match, Specific match) and the generation of the imitation.

Matching accuracy for all question types was above chance for the first generation of imitations, b = 1.65 (SE = 0.14) log-odds, odds = 0.50, z = 11.58, $p <$ 0.001, and decreased steadily over generations, b = -0.16 (SE = 0.04) log-odds, z = -3.72, $p <$ 0.001. We tested whether this increase in difficulty was constant across the three types of questions or if some question types became more difficult than others. The results are shown in Fig. 4B. Performance decreased over generations more rapidly for questions requiring a within-category distinction than for between-category questions, b = -0.08 (SE = 0.03) log-odds, z = -2.68, p = 0.007, suggesting that between-category information was more resistant to loss through repeated imitation. An alternative explanation for this result is that the within-category match questions are simply more difficult[2] because the sounds provided as choices are more acoustically similar to one another than the between-category questions, and therefore, performance might be expected to drop off more rapidly with repeated imitations. However, performance also decreased for the easiest type of question where the correct answer was the actual seed generating the imitation (True seed questions; see Fig. 4A); the advantage of having the true seed among between-category distractors decreased over generations, b = -0.07 (SE = 0.02) log-odds, z = -2.77, p = 0.006. The observed increase in the "category advantage" (i.e., the advantage of having between-category distractors) combined with a decrease in the "true seed advantage" (the advantage of having the actual seed among the choices), shows that the changes induced by repeated imitation caused the imitations to lose some of properties that linked the earlier imitations to the specific sound that motivated them, while nevertheless preserving a more abstract category-based resemblance.

We next tested whether it was possible to match the written transcriptions of the auditory sounds back to the original environmental sounds. Participants were given a novel word (the most frequent transcriptions of first and last generation imitations) and had to guess the sound that was represented by the invented word. The distractors for all questions were between-category, i.e. true seed and category match. Specific match questions were omitted.

Remarkably, participants were able to guess the correct meaning of a word that was transcribed from an imitation that had been repeated up to 8 times, b = 0.83 (SE = 0.13) log-odds, odds = -0.18, z = 6.46, $p <$ 0.001 (Fig. 4C). This was true for True seed questions containing the actual seed generating the transcribed imitation, b = 0.75 (SE = 0.15) log-odds, z = 4.87, $p <$ 0.001, and for Category match questions where participants had to associate transcriptions with a particular category of environmental sounds, b = 1.02 (SE = 0.16) log-odds, z = 6.39, $p <$ 0.001. The effect of generation did not vary across these question types, b = 0.05 (SE = 0.10) log-odds, z = 0.47, p = 0.638. Possible reasons for this difference between imitations and their transcriptions are explored in the Supporting Information.

In sum, our results show how unguided repetition causes initial imitations of environmental sounds transition to more word-like forms. They suggest that in the course of this transition, the imitations become more categorical and more effective as learned category labels all while retaining some resemblance to the environmental sounds that motivated them.

Discussion

Imitative (or "iconic") words are found across the spoken languages of the world (Dingemanse et al., 2015; Imai & Kita, 2014; Perniss et al., 2010). Counter to past assumptions about the limitations of human vocal imitation, people are surprisingly effective at using vocal imitation to represent and communicate about the sounds in their environment (Lemaitre et al., 2016) and more abstract meanings (Perlman et al., 2015), making the hypothesis that early spoken words originated from imitations a plausible one. We examined whether simply repeating an imitation of an environmental sound—with no intention to create a new word or even to communicate—produces more word-like forms.

[2] We observed that performance on some Specific match questions dropped below chance for later generations indicating participants had an apparent aversion to the nominally correct answer. Additional analyses showed that participants were not converging on a single incorrect response. The reason for this pattern is at present unclear. Removing these trials from the analysis does not substantively change the conclusions.

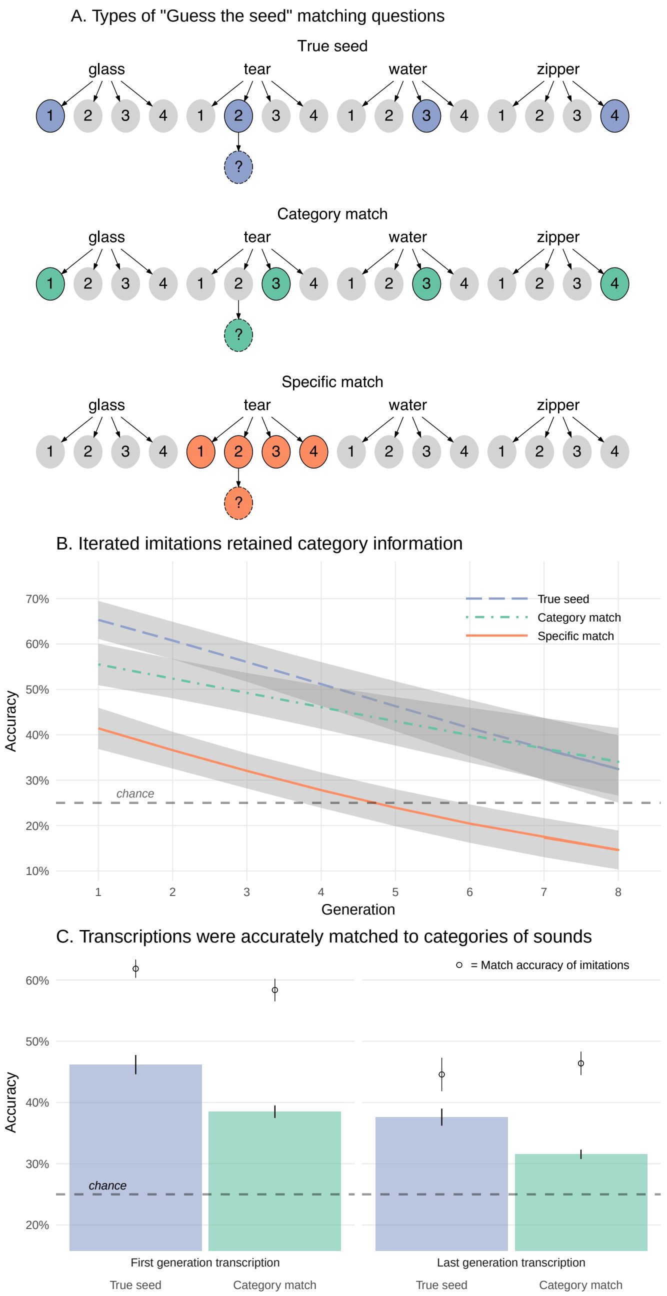


Figure 4: A. Three types of matching questions used to assess the resemblance between the imitation (and transcriptions of imitations) and the original seed sounds. For each question, participants listened to an imitation (dashed circles) and had to guess which of 4 sound choices (solid circles) they thought the person was trying to imitate. True seed questions contained the actual sound that generated the imitation as one of the choices (correct response). The remaining sounds were sampled from different categories. Category match questions replaced the original seed sound with another sound from the same category. Specific match questions pitted the actual seed against the other seeds within the same category. B. Change in matching accuracy over generations of imitations, shown as predictions of the generalized linear models with ±1 SE of the model predictions. The "category advantage" (Category match vs. Specific match) increased over generations, while the "true seed advantage" (True seed v. Category match) decreased (see main text), suggesting that imitations lose within-category information more rapidly than between-category information. C. Change in matching accuracy over generations of imitations transcribed into English-sounding words. Imitations and transcriptions of imitations could still be matched back to the category of sound that motivated the original imitation even after 8 generations. Match accuracy for imitations is shown for comparison.

Our results show that through simple repetition, imitative vocalizations became more word-like both in form and function. In form, the vocalizations gradually stabilized over generations, becoming more similar from imitation to imitation. They also became increasingly standardized in accordance with English orthography, as later generations were more consistently transcribed into English words. In function, the increasingly word-like forms became more effective as category labels. In a category learning experiment, naïve participants were faster to learn category labels derived from transcriptions of later-generation imitations than those derived from direct imitations of the environmental sound. This fits with previous research showing that the relatively arbitrary forms that are typical of words (e.g. "dog") makes them better suited to function as category labels compared to direct auditory cues (Boutonnet & Lupyan, 2015; Edmiston & Lupyan, 2015; e.g. the sound of a dog bark; Lupyan & Thompson-Schill, 2012).

Evan as the vocalizations became more word-like, they nevertheless maintained an imitative quality. After eight generations they could no longer be matched to the particular sound from which they originated any more accurately than they could be matched to the general category of environmental sound. Thus, information that distinguished an imitation from other sound categories was more resilient to transmission decay than exemplar information within a category. Remarkably, even after the vocalizations were transcribed into English orthography, participants were able to guess their original sound category from the written "word". In contrast to the vocalizations, participants continued to be more accurate at matching late generation transcriptions back to their particular source sound relative to other exemplars from the same category.

Although the number of imitative words in contemporary languages may appear to be very small (Crystal, 1987; Newmeyer, 1992), increasing evidence from disparate languages shows that vocal imitation is, in fact, a widespread source of vocabulary. Cross-linguistic surveys indicate that onomatopoeia—imitative words used to represent sounds—are a universal lexical category found across the world's languages (Dingemanse, 2012). Even English, a language that has been characterized as relatively limited in iconic vocabulary (Vigliocco, Perniss, & Vinson, 2014), is documented to have hundreds of clearly imitative words including words for human and animal vocalizations as well as various types of environmental sounds (Rhodes, 1994; Sobkowiak, 1990). Besides words that are directly imitative of sounds—the focus of the present study—many languages contain semantically broader inventories of ideophones. These words comprise a grammatically and phonologically distinct class of words that are used to express various sensory-rich meanings, such as qualities related to manner of motion, visual properties, textures and touch, inner feelings and cognitive states (Dingemanse, 2012; Nuckolls, 1999; Voeltz & Kilian-Hatz, 2001). As with onomatopoeia, ideophones are often recognized by naïve speakers as bearing a degree of resemblance to their meaning (Dingemanse, Schuerman, & Reinisch, 2016).

Our study focused on imitations of environmental sounds and more work remains to be done to determine the extent to which vocal imitation can ground de novo vocabulary creation in other semantic domains (Lupyan & Perlman, 2015; e.g., Perlman et al., 2015). What the present results make clear is that the transition from imitation to word can be a rapid and simple process: the mere act of repeated imitation can drive vocalizations to become more word-like in both form and function. Notably, just as onomatopoeia and ideophones of natural languages maintain a resemblance to the quality they represent, the present vocal imitations transitioned to words while retaining a resemblance to the original sound that motivated them.

Methods

Selecting seed sounds

To avoid sounds having lexicalized or conventionalized onomatopoeic forms in English, we used inanimate categories of environmental sounds. Using an odd-one-out norming procedure (N=105 participants; see Supporting Information), an initial set of 36 sounds in 6 categories was reduced to a final set of 16 "seed" sounds: 4 sounds in each of 4 categories (Figs. 5-6). The four final categories were: water, glass, tear, zipper.

Collecting imitations

Participants (N=94) recruited from Amazon Mechanical Turk were paid to participate in an online version of the children's game of "Telephone". Participants were instructed that they would hear some sound and their task is to reproduce it as accurately as possible using their computer microphone (Fig. 7). Full instructions are provided in the Supporting Information. Participants listened to and imitated 4 sounds, receiving one sound from each of the four categories of sounds drawn at random such that participants were unlikely to hear the same person more than once. Recordings that were too quiet (less than –30 dBFS) were not allowed. Imitations were monitored by an experimenter to catch any gross errors in recording before they were heard by the next generation of imitators (Fig. 8). For example, recordings were trimmed to the length of the imitation, and recordings with loud sounds in the background were removed. The experimenter also blocked sounds that violated the rules of the experiment, e.g., by saying something in English. A total of 115 (24%) imitations were removed prior to subsequent analysis.

Measuring acoustic similarity

Acoustic similarity was measured by having research assistants listen to pairs (approx. 314) of sounds and rate their subjective similarity. On each trial, raters heard two sounds from subsequent generations were played in succession but in random order. They then indicated the similarity between the sounds on a 7-point Likert scale from *Entirely different and would never be confused* to *Nearly identical*. Raters were encouraged to use as much of the scale as they could while maximizing the likelihood that, if they did this procedure again, they would reach the same judgments. Full instructions are provided in the Supporting Information. Ratings were normalized (z-scored) prior to analysis.

Collecting transcriptions of imitations

Participants (N=216) recruited from Amazon Mechanical Turk were paid to transcribe sounds into words in an online survey. They listened to imitations and were instructed to write down what they heard as a single word so that the written word would sound as much like the message as possible. Exact instructions are provided in the Supporting Information (Fig. 9).

Transcriptions were generated from the first and last three generations of all imitations collected in the Telephone game; that is, not all imitations were transcribed (Fig. 10). Participants also provided transcriptions of the original environmental seed sounds (Fig. 11). Transcriptions from participants who failed a catch trial were excluded (N=2), leaving 2163 transcriptions for analysis. Of these, 179 transcriptions (8%) were removed because they contained English words, which was a violation of the instructions of the experiment.

Learning transcriptions as category labels

Our transmission chain design and subsequent transcription procedure created 1814 unique words. From these, we sampled words transcribed from first and last generation imitations as well as from seed sounds that were equated in length and overall matching accuracy. Specifically, we removed transcriptions that contained less than 3 unique characters and transcriptions that were over 10 characters long. Of the remaining transcriptions, a sample of 56 were selected using a bootstrapping procedure to have approximately equal means and variances of overall matching accuracy. The full procedure for sampling the words in this experiment is described in the Supporting Information.

Participants (N=67) were University of Wisconsin undergraduates who received course credit for participation. Participants were randomly assigned four novel labels to learn for four categories of environmental sounds. Participants were assigned between-subject to learn labels (transcriptions) of the first or last generation imitations, as well as labels from transcriptions of seed sounds as a control (Fig. 15). On each trial, participants heard one of the 16 seed sounds. After a 1s delay , participants saw a label–one of the transcribed imitations–and responded yes or no using a gamepad controller depending on whether the sound and the word went together. Participants received accuracy feedback (a bell if correct; a buzzing sound if incorrect). Four outlier participants were excluded from the final sample due to high error rates and slow RTs.

Participants categorized all 16 seed sounds over the course of the experiment, but they learned them in blocks of 4 sounds at a time. Within each block of 24 trials, participants heard the same four sounds and the same four words multiple times, with a 50% probability of the sound matching the word on any given trial. At the start of a new block of trials, participants heard four new sounds they had not heard before, and had to learn to associate these new sounds with the words they had learned in the previous blocks.

Matching imitations to seed sounds

Participants (N=751) recruited from Amazon Mechanical Turk were paid to listen to imitations, one at a time, and for each one, choose one of four possible sounds they thought the person was trying to imitate. The task was unspeeded and no feedback was provided. Participants completed 10 questions at a time.

Question types (True seed, Category match, Specific match) were assigned between-subject. Participants in the True seed and Category match conditions were provided four seed sounds from different categories as choices in each question. Participants in the Specific match condition were provided four seed sounds from the same category. All 365 imitations were tested in each of the three conditions.

Matching transcriptions to seeds

Participants (N=468) recruited from Amazon Mechanical Turk completed a modified version of the matching survey. Instead of listening to imitations, participants now read a word (a transcription of an imitation), which they were told was an invented word. They were instructed that the word was invented to describe one of the four presented sounds, and they had to guess which one. Of all the unique transcriptions that were collected for each sound (imitations and seed sounds), only the top four most frequent transcriptions were used in the matching experiment. 6 participants failed a catch trial and were excluded, leaving 461 participants in the final sample.

Supplementary Materials

Open data and materials

We are committed to making the results of this research open and reproducible. The R code used to generate all stats and figures reported in the main manuscript as well as in this Supporting Information document is available on GitHub at github.com/lupyanlab/creating-words. The data are available in an R package, which can be downloaded and installed with the following R commands:

```
# Install the R package from GitHub
library(devtools)
install_github("lupyanlab/words-in-transition",
               subdir = "wordsintransition")

# Load the package
library(wordsintransition)

# Browse all datasets
data(package = "wordsintransition")
```

```
# Load a particular dataset
data("acoustic_similarity_judgments")
```

The materials used to run the experiments are also available in GitHub repositories. The web app used to collect vocal imitations, transcriptions of imitations, and matches of imitations and transcriptions to the original seed sounds is available at github.com/lupyanlab/telephone-app. Analyses of acoustic similarity including both algorithmic analyses as well as the procedure for gathering subjective judgments of similarity are provided at github.com/lupyanlab/acoustic-similarity. The category learning experiment is available at github.com/lupyanlab/learning-sound-names.

Selecting seed sounds

Our goal in selecting sounds to serve as seeds for the transmission chains was to pick multiple sounds within a few different categories such that each category member was approximately equally distinguishable from the other sounds within the same category. To do this, we started with an initial set of 6 categories and 6 sounds in each category and conducted 2 rounds of “odd one out” norming to reduce the initial set to a final set of 16 seed sounds: 4 sounds in each of 4 categories. Having 4 sounds in 4 categories was the minimum necessary in order to generate 4AFC questions with both between-category and within-category distractors with the appropriate level of counterbalancing across all conditions.

Participants (N=105) recruited via Amazon Mechanical Turk were paid to participate in the norming procedure. Participants listened to all sounds in each category and picked the one that they thought was **the most different** from the others. In the first round of norming, participants listened to 6 sounds on a given trial. We removed the 2 sounds in each category that were the most different from the others (Fig. 5), and repeated the norming process again with 4 sounds in each category (Fig. 6). After the second round of norming, we selected the four categories to use in the experiment. The resulting sounds that were selected in each category are considered to be a set of equally distinguishable category members.

The final 16 seed sounds used in the transmission chain experiment can be downloaded from sapir.psych.wisc.edu/telephone/seeds/all-seeds.zip.

Collecting vocal imitations

Participants played a version of the children’s game of Telephone via a web-based interface (Fig. 7). Initially the only action available to participants is to play the message by clicking the top sound icon. After listening to the message once, they could then initiate a recording of their imitation by clicking the bottom sound icon to turn the recorder on. Turning the recorder off submitted their response. If the recording was too quiet (less than –30 dBFS), participants were asked to repeat their imitation. In response, they could repeat the initial message again. After a successful recording was submitted, a new message was loaded. Participants made 4 recordings each. The instructions given to participants are presented below.

> We are researchers at the University of Wisconsin-Madison studying how audio messages are passed on from person to person, much like in the children’s game Telephone. If you choose to participate, we will ask you to listen to an audio message recorded by someone else, and then record yourself imitating the message that you heard using your computer’s microphone.
>
> Unlike the children’s game of Telephone, the sounds you will hear will not be recognizable English words, but will be various nonspeech sounds. Your task is the same, however: to recreate the sound you heard as accurately as you can. [. . .]

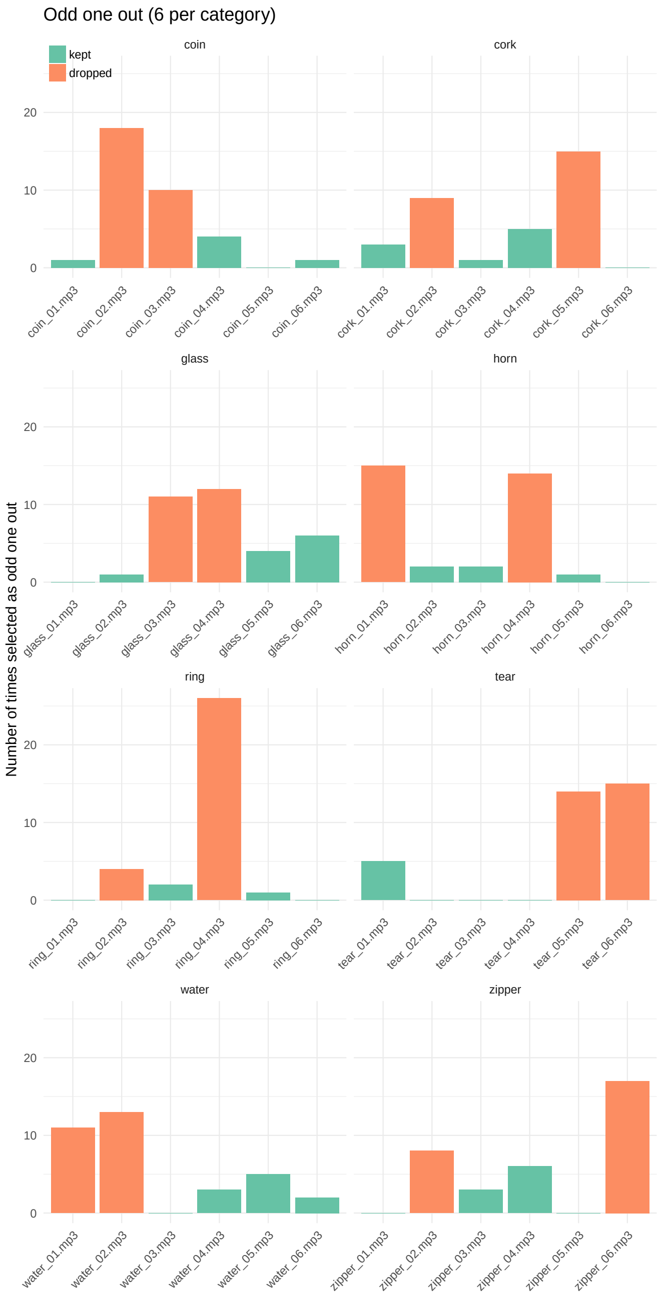


Figure 5: Results of the first round of seed norming. After collecting these data, two sounds were removed from each category and the norming procedure was conducted again.

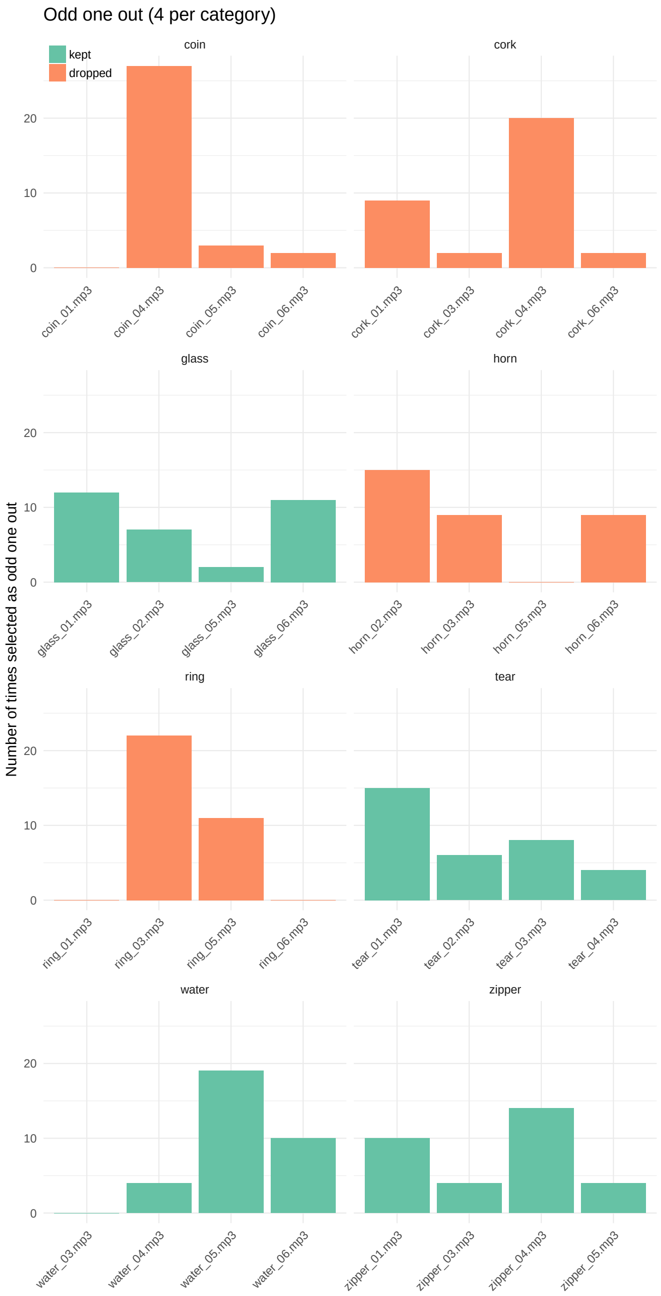


Figure 6: Results of the second round of seed norming. After collecting these data, four categories of sounds were selected to use in the main experiment.

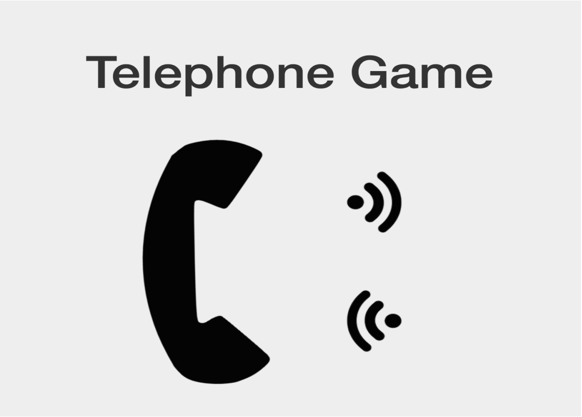


Figure 7: The interface for collecting vocal imitations. Participants clicked the top sound icon to hear the message and the bottom sound icon to record their response. After ending their recording a new message was presented.

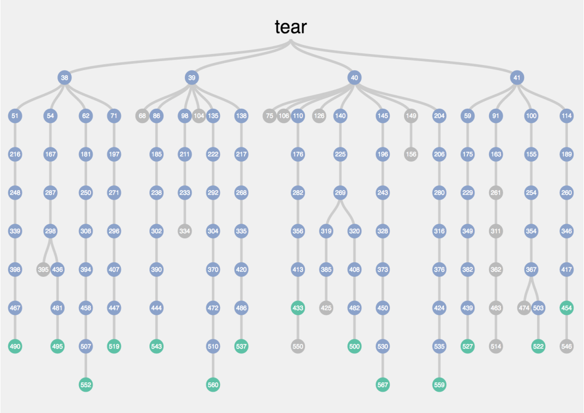


Figure 8: The interface for monitoring incoming imitations. All imitations were listened to by an experimenter and trimmed to remove extraneous noise. Imitations eligible for the next generation appear in green. Bad quality imitations were rejected (in gray).

Monitoring incoming imitations

Since the imitations were collected online, it was likely that at least some of the imitations would be invalid, either due to low recording quality or due to a violation of the instructions of the experiment (e.g., saying something in English). We monitored the imitations as they were received to verify the integrity of the recordings and exclude ones where necessary. The monitoring helped catch gross errors in the timing of the recording, the most common of which was recordings that were too long relative to the imitation. Via this interface (Fig. 8), recordings were heard, trimmed, and, in some cases, rejected. Due to random assignment and the irregular nature of the rejections, all transmissions chains did not reach to the full 8 generations.

Measuring acoustic similarity

After collecting the imitations in the transmission chain design, the imitations were submitted to analyses of acoustic similarity. The primary measure of acoustic similarity was obtained from research assistants who participated in a randomized rating procedure. We also measured algorthmic acoustic distance.

Acoustic similarity judgments

Five research assistants rated the similarity between 324 different pairs of imitations. These pairs comprised consecutive imitations in the transmission chain design, e.g., each message was compared to its response. Message order was randomized on each trial so that participants did not know which message was the original and which message was the imitation. Participants were also blind to the overall generation of the imitations by randomizing generation from trial to trial. To facilitate consistency in rating, pairs of sounds were blocked by category, e.g., participants rated all tearing sounds before moving on to other categories of sounds. The instructions given to participants are stated below.

> On each trial, you will hear two sounds played in succession. To help you distinguish them, during the first you will see the number 1, and during the second a number 2. After hearing the second sound, you will be asked to rate how similar the two sounds are on a 7-point scale.
>
> A 7 means the sounds are nearly identical. That is, if you were to hear these two sounds played again, you would likely be unable to tell whether they were in the same or different order as the first time you heard them. A 1 on the scale means the sounds are entirely different and you would never confuse them. Each sound in the pair will come from a different speaker, so try to ignore differences due to just people having different voices. For example, a man and a woman saying the same word should get a high rating.
>
> Please try to use as much of the scale as you can while maximizing the likelihood that if you did this again, you would reach the same judgments. [...]

Algorithimic measures of acoustic similarity

To obtain algorithmic measures of acoustic similarity, we used the acoustic distance functions included in the Phonological Corpus Tools program (Hall, Allen, Fry, Mackie, & McAuliffe, n.d.). Using this program, we computed MFCC similarities between pairs of sounds using 12 coefficients in order to obtain speaker-independent estimates.

We calculated average acoustic similarity in six kinds of comparisons (Fig. 9A). The first four kinds compared imitations within the same category of environmental sound (glass, tear, water, zipper). The most similar were imitations along consecutive transmissions chains (Within chain, consecutive). Next were all pairwise comparisons of imitations from the same chain (Within chain), followed by all pairwise comparisons leading from the same seed sound (Within seed), and finally all pairwise comparisons for imitations from all seeds within the same category (Within category). As expected, all four kinds of within category comparisons resulted in higher similarity scores than the between category comparisons. The between category comparisons included imitations from the same generation across different chains (Between category, same), and imitations from consecutive generations from different chains (Between category, consecutive).

In parallel with the judgments of acoustic similarity, we also investigated how automated measures of acoustic similarity change over generations of imitation. For the automated analyses we did not find a reliable relationship between imitation generation and automated analysis of acoustic similarity, $b = 0.04$ (SE = 0.03), $t(357.0) = 1.18$, $p = 0.24$ (Fig. 10B). For our stimuli the correlation between automated analyses of acoustic similarity and rater judgments was low, $r = 0.20$, 95% CI [0.16, 0.25] (Fig. 10C), suggesting that the automated analyses may not capture the acoustic features driving the perception of acoustic similarity of these stimuli. This is possibly due to the non-verbal nature of the imitations as well as variation in recording quality between participants in the online study.

Collecting transcriptions of imitations

To collect transcriptions of vocal imitations, participants were instructed to turn the sound they heard into a word that, when read, would sound much like the imitation. The interface for collecting transcriptions as well

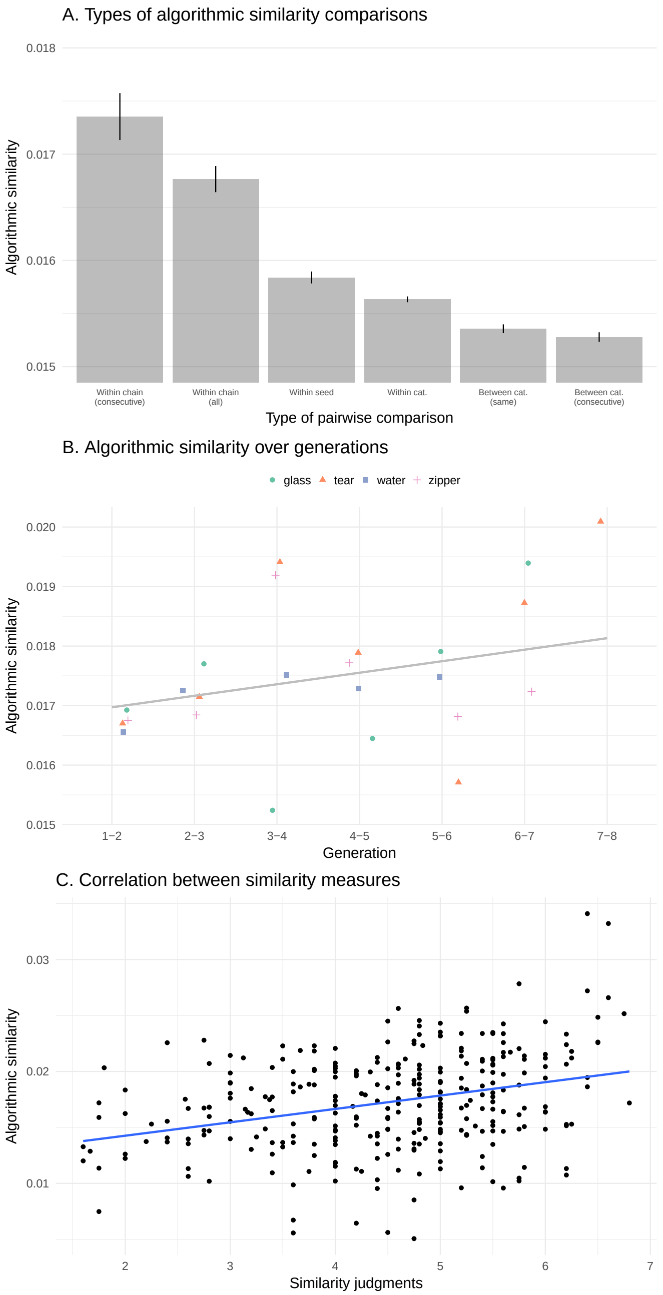
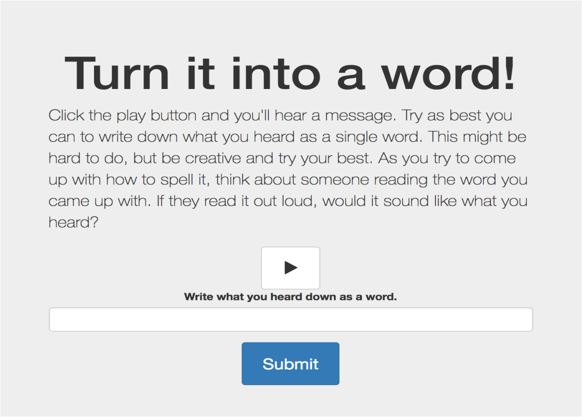


Figure 9: Algorithmic measures of acoustic distance. A. Average acoustic distance between pairs of sounds grouped by type of comparison. B. Change in algorithmic acoustic distance over generations of imitations. C. Correlation between similarity judgments and algorithmic measures.



Turn it into a word!

Click the play button and you'll hear a message. Try as best you can to write down what you heard as a single word. This might be hard to do, but be creative and try your best. As you try to come up with how to spell it, think about someone reading the word you came up with. If they read it out loud, would it sound like what you heard?

Figure 10: Interface for collecting transcriptions. Participants listened to an imitation and were instructed to create novel words corresponding to the sound they heard.

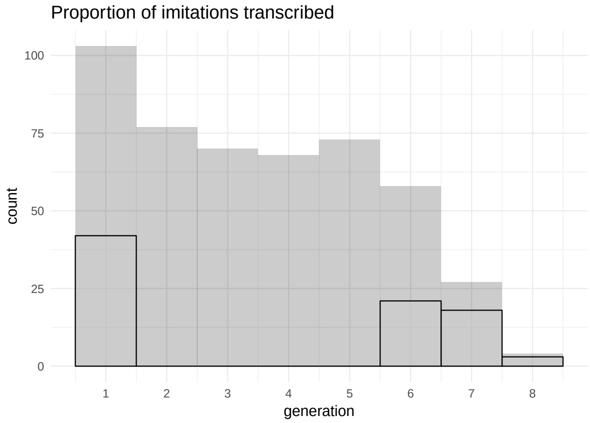


Figure 11: Proportion of imitations that were transcribed. Gray region indicates the number of imitations collected at each generation. Outlined regions denote the number of imitations that were transcribed. First generation imitations and the last three generations of imitations were transcribed.

as the exact wording of the instructions is shown in Fig. 10.

We only obtained transcriptions of a sample of the imitations collected in the Telephone game. Specifically, we obtained transcriptions of the first generation of imitations as well as the last 3 generations. The proportion of imitations that were transcribed is shown in Fig. 11.

Alternative measures of orthographic distance

Our primary measure of transcription difference was provided by the `SequenceMatcher` functions in the `difflib` package of the python standard library. These functions implement Ratcliff and Obershelp's "gestalt pattern matching" algorithm, with the additional feature of taking into account repeated "junk" characters when finding longest contiguous substring matches. Here we report alternative measures of orthographic distance, such as the number of exact spelling matches (Fig. 12A).

As can be seen in Fig. 12A, some of the imitations did not yield any exact string matches, indicating that all transcriptions for these imitations were unique. This potentially invalidates our metric for measuring average distance since it involved comparing the most frequent transcription to all other transcriptions of a given imitation. For imitations with all unique transcriptions, the "most frequent" transcription was selected at random. In Fig. 12B, we show the results of our orthographic distance metric separately for imitations with and without any agreement.

Fig. 12C shows an alternative measure corresponding exactly to the length of the substring match among transcriptions, again separating the results by whether or not there was any agreement on the transcription of the imitation.

Matching imitations and transcriptions to seeds

To measure the extent to which imitations resembled their seed sound source, we tasked participants with matching the imitation (Fig. 13) or a transcription of an imitation (Fig. 14) to its source relative to other seed sounds used in the experiment. Participants were assigned 4 seed sounds (between-subject) to serve as options in the 4AFC task. Mousing over the options played the sounds, which became active after the participant listened to the imitation one time completely. For imitations, they were allowed to listen to the imitation as many times as they wanted. On each trial they were presented a different imitation and asked to match it to the seed sound they thought the imitator was trying to imitate. For transcriptions, they were instructed the that word was "invented" to correspond to one of the sounds in their options.

Learning transcriptions as category labels

To determine which transcriptions to test as category labels, we first selected only those transcriptions which had above chance matching performance when matching back to the original seeds. (The matching experiments were conducted chronologically prior to the category learning experiment). Then we excluded transcriptions that had less than two unique characters or were over 10 characters long, and used a bootstrapping procedure to sample from both first and last generation imitations to reach a final set that controlled for overall matching accuracy. The R script that performed the selection and bootstrapping procedure is available on GitHub at github.com/lupyanlab/learning-sound-names/blob/master/R/select_messages.R. It involves selecting a desired mean matching accuracy from the last generation of transcriptions, and sampling transcriptions from first generation transcriptions until the sample falls within the desired variance.

In the experiment, participants learned, through trial-and-error, the names for four different categories of sounds. On each trial participants listened to one of the 16 environmental sounds used as seeds and then saw a novel word–a transcription of one of the imitations. Participants responded by pressing a green button on a gamepad controller if the label was the correct label and a red button otherwise. They received accuracy feedback after each trial.

The experiment was divided into blocks so that participants had repeated exposure to each sound and the novel labels multiple times within a block. At the start of a new block, participants received four new sounds from the same four categories (e.g., a new zipping sound, a new water-splash sound, etc.) that they had not heard before, and had to associate these sounds with the same novel labels from the previous blocks. The extent to which their performance declined at the start of each block serves as a measure of how well the label they associated with the sound worked as a label for the category.

Transcriptions of seed sounds

As a control, we also had participants generate "transcriptions" directly from the seed sounds. These transcriptions were the most variable in terms of spelling (Fig. 16A), but the most frequent of them were the easiest to match back to the original seeds (Fig. 16C). When learning these transcriptions as category labels, participants were the fastest to learn them in the first block (Fig. 16B), but they did not generalize to new category members as fast as transcriptions taken from last generation imitations.

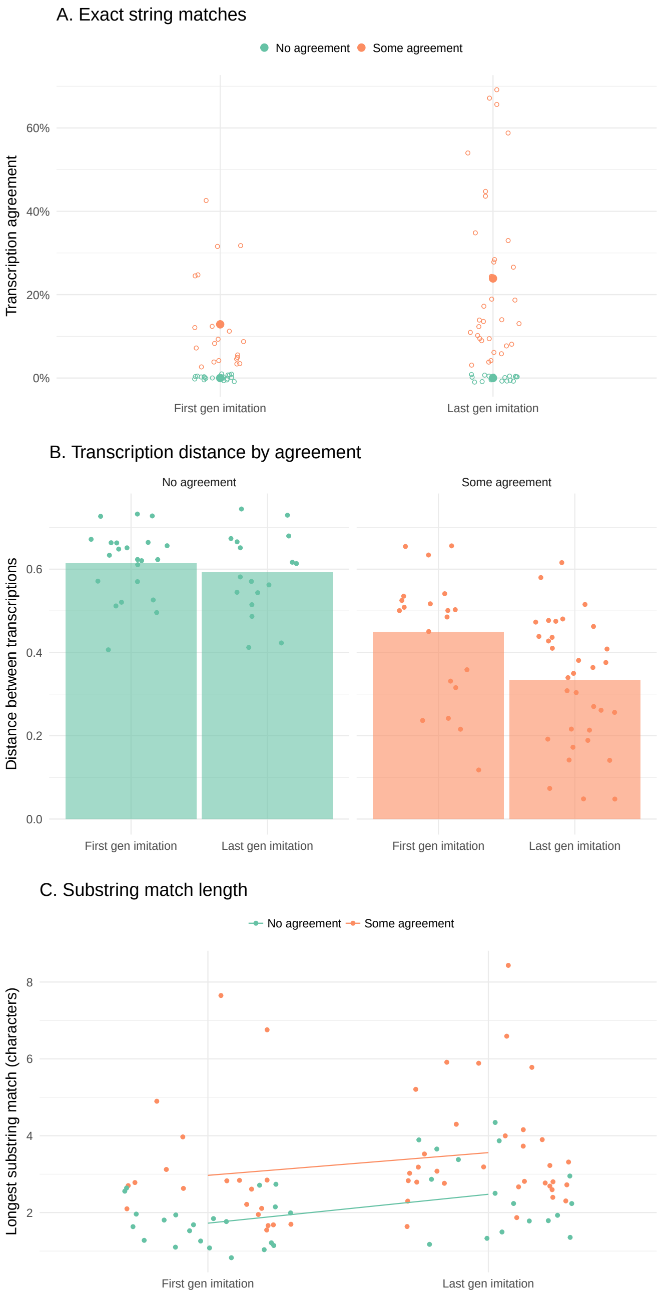
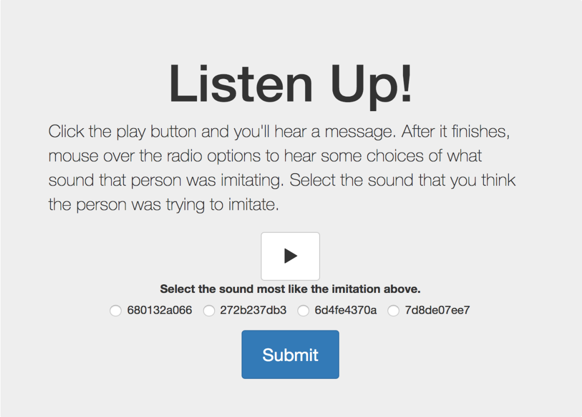


Figure 12: Alternative measures of orthographic distance. A. Percentage of exact string matches per imitation. B. Orthographic distance separated by whether there was any agreement among the transcriptions of a given imitation. C. Change in the average length of the substring match.



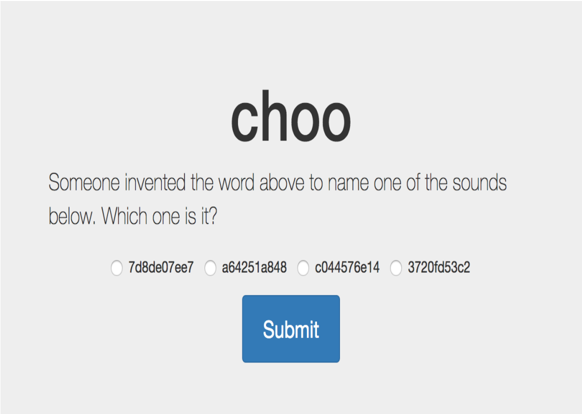
Listen Up!

Click the play button and you'll hear a message. After it finishes, mouse over the radio options to hear some choices of what sound that person was imitating. Select the sound that you think the person was trying to imitate.

Select the sound most like the imitation above.

○ 680132a066 ○ 272b237db3 ○ 6d4fe4370a ○ 7d8de07ee7

Figure 13: Interface for matching imitations back to original seed sounds.



choo

Someone invented the word above to name one of the sounds below. Which one is it?

○ 7d8de07ee7 ○ a64251a848 ○ c044576e14 ○ 3720fd53c2

Figure 14: Interface for matching transcriptions back to original seed sounds.

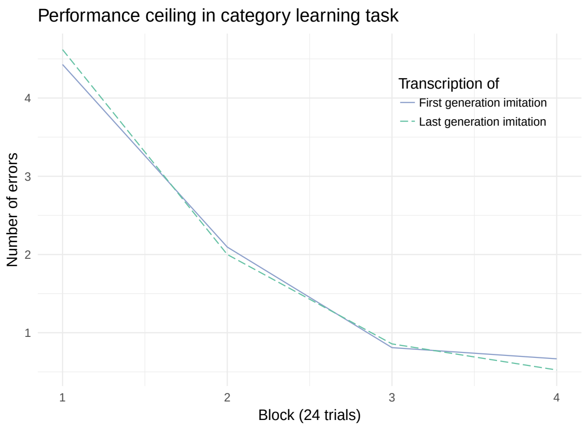
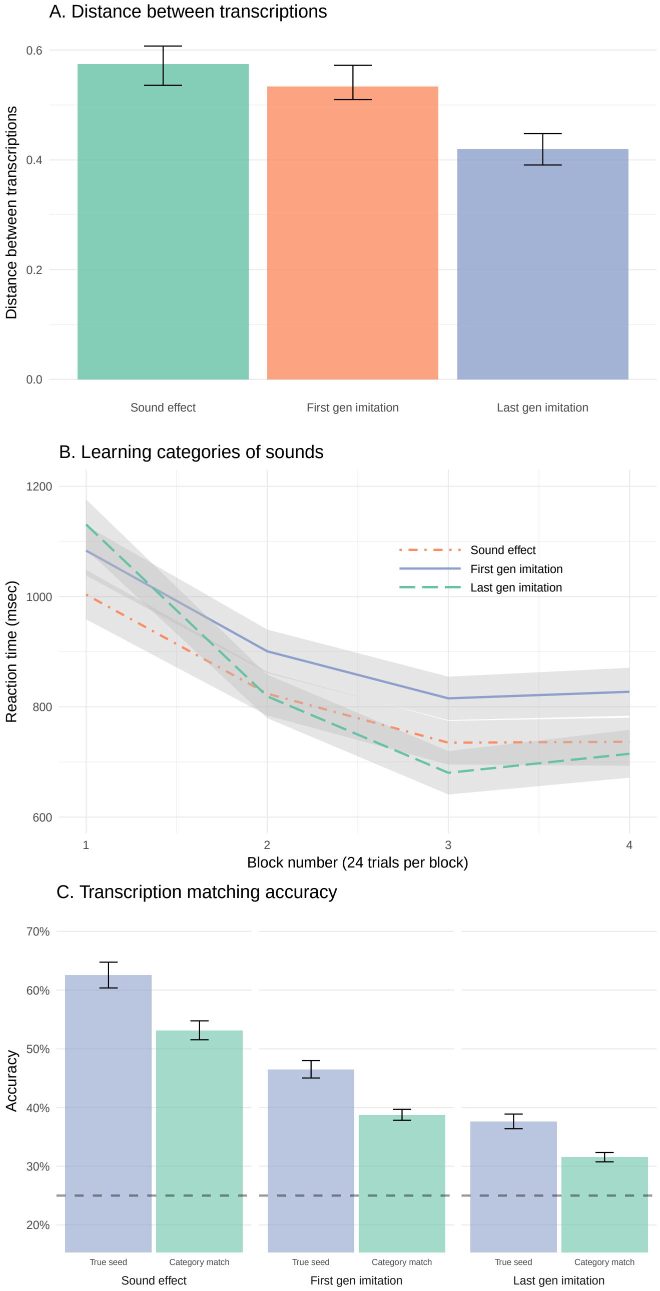


Figure 15: Mean number of errors by block of 24 trials showing that accuracy performance quickly reached ceiling after the first block of trials.



Acknowledgements

The authors acknowledge Jesse Reid, Hailey Schiedermayer, Zoe Hansen, Maggie Parker, and Yacong Wu for conducting the acoustic similarity analysis.

References

Arbib, M. A. (2012). *How the brain got language: The mirror system hypothesis* (Vol. 16). Oxford University

Press.

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

Boutonnet, B., & Lupyan, G. (2015). Words Jump-Start Vision: A Label Advantage in Object Recognition. *Journal of Neuroscience*, *35*(25), 9329–9335.

Brown, R. W., Black, A. H., & Horowitz, A. E. (1955). Phonetic symbolism in natural languages. *Journal of Abnormal Psychology*, *50*(3), 388–393.

Clark, H. H., & Gerrig, R. J. (1990). Quotations as demonstrations. *Language*, *66*, 764–805.

Corballis, M. C. (2003). *From hand to mouth: The origins of language*. Princeton University Press.

Crystal, D. (1987). *The Cambridge Encyclopedia of Language* (Vol. 2). Cambridge Univ Press.

Dingemanse, M. (2012). Advances in the Cross-Linguistic Study of Ideophones. *Language and Linguistics Compass*, *6*(10), 654–672.

Dingemanse, M. (2014). Making new ideophones in Siwu: Creative depiction in conversation. *Pragmatics and Society*.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, Iconicity, and Systematicity in Language. *Trends in Cognitive Sciences*, *19*(10), 603–615.

Dingemanse, M., Schuerman, W., & Reinisch, E. (2016). What sound symbolism can and cannot do: Testing the iconicity of ideophones from five languages. *Language*, *92*.

Donald, M. (2016). Key cognitive preconditions for the evolution of language. *Psychonomic Bulletin & Review*, 1–5.

Edmiston, P., & Lupyan, G. (2015). What makes words special? Words as unmotivated cues. *Cognition*, *143*(C), 93–100.

Goldin-Meadow, S. (2016). What the hands can tell us about language emergence. *Psychonomic Bulletin & Review*, *24*(1), 1–6.

Hall, K. C., Allen, B., Fry, M., Mackie, S., & McAuliffe, M. (n.d.). Phonological CorpusTools. *14th Conference for Laboratory Phonology*.

Hewes, G. W. (1973). Primate Communication and the Gestural Origin of Language. *Current Anthropology*, *14*(1/2), 5–24.

Hockett, C. F. (1978). In search of Jove's brow. *American Speech*, *53*(4), 243–313.

Imai, M., & Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651).

Kendon, A. (2014). Semiotic diversity in utterance production and the concept of 'language'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130293–20130293.

Klima, E. S., & Bellugi, U. (1980). *The signs of language*. Harvard University Press.

Kuznetsova, A., Bruun Brockhoff, P., & Haubo Bojesen Christensen, R. (2016). *lmerTest: Tests in Linear Mixed Effects Models*.

Lemaitre, G., & Rocchesso, D. (2014). On the effectiveness of vocal imitations and verbal descriptions of sounds. *The Journal of the Acoustical Society of America*, *135*(2), 862–873.

Lemaitre, G., Houix, O., Voisin, F., Misdariis, N., & Susini, P. (2016). Vocal Imitations of Non-Vocal Sounds.

PloS One, *11*(12), e0168167–28.

Lewis, J. (2009). As well as words: Congo Pygmy hunting, mimicry, and play. In *The cradle of language*. The cradle of language.

Locke, J. (1948). An essay concerning human understanding. In W. Dennis (Ed.), *Readings in the history of psychology*. Norwalk, CT.

Lupyan, G., & Perlman, M. (2015). The vocal iconicity challenge! In *The th biennial protolanguage conference*. Rome, Italy.

Lupyan, G., & Thompson-Schill, S. L. (2012). The evocative power of words: Activation of concepts by verbal and nonverbal means. *Journal of Experimental Psychology: General*, *141*(1), 170–186.

Newmeyer, F. J. (1992). Iconicity and generative grammar. *Language*.

Nuckolls, J. B. (1999). The case for sound symbolism. *Annual Review of Anthropology*, *28*(1), 225–252.

Perlman, M., Dale, R., & Lupyan, G. (2015). Iconicity can ground the creation of vocal symbols. *Royal Society Open Science*, *2*(8), 150152–16.

Perniss, P., Thompson, R. L., & Vigliocco, G. (2010). Iconicity as a General Property of Language: Evidence from Spoken and Signed Languages. *Frontiers in Psychology*, *1*.

Pinker, S., & Jackendoff, R. (2005). The faculty of language: what's special about it? *Cognition*, *95*(2), 201–236.

Plato, & Reeve, C. D. C. (1999). *Cratylus*. Indianapolis: Hackett.

Rhodes, R. (1994). Aural images. *Sound Symbolism*, 276–292.

Sobkowiak, W. (1990). On the phonostatistics of English onomatopoeia. *Studia Anglica Posnaniensia*, *23*, 15–30.

Tomasello, M. (2010). *Origins of human communication*. MIT press.

Vigliocco, G., Perniss, P., & Vinson, D. (2014). Language as a multimodal phenomenon: implications for language learning, processing and evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130292–20130292.

Voeltz, F. E., & Kilian-Hatz, C. (2001). *Ideophones* (Vol. 44). John Benjamins Publishing.