The emergence of words from vocal imitations

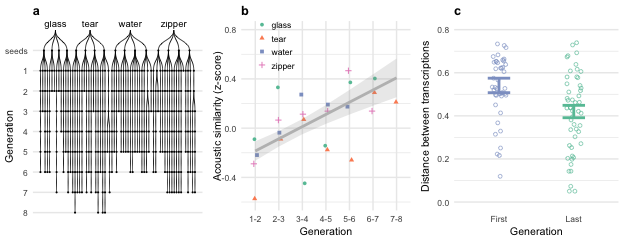
Pierce Edmiston, Marcus Perlman, and Gary Lupyan

**People have long pondered the origins of language, especially the words that compose them1,2. Here, we report a large-scale experiment (*N*=1571) investigating how conventional spoken words might emerge from imitations of environmental sounds. Does the repeated imitation of an environmental sound gradually give rise to novel word forms? In what ways do these words resemble the original sounds that motivated them? 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 for a maximum of 8 generations. The results showed that the imitations became more stable and word-like, and later imitations were easier to learn 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 while continuing to retain a resemblance to the original sound that motivated them, and speak to the possible role of human vocal imitation in explaining the origins of at least some spoken words.**

The importance of imitation and depiction in the origin of signs is clearly observable in signed languages3–5, but in considering the idea that imitation in the vocal modality 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 role6–11. For example, Pinker and Jackendoff12 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 languages13,14 and the frequent use of vocal imitation and depiction in spoken discourse15,16. This has led some to argue for the importance of imitation for understanding the origin of spoken words17–21. 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 words22. 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 listeners20. 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 listeners23. 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 decay24.

Thus, converging evidence suggests that people can use vocal imitation as an effective means of communication. But can vocal imitations ever give rise to words that can be integrated into the vocabulary of a language? And if so, 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. The final set of vocal imitations included 365 imitations along 105 contiguous transmission chains from 94 participants (Fig. 1a).



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. Change in perception of acoustic similarity over generations of repetition. Mean acoustic similarity ratings of imitations in each category are shown, along with the predictions of a linear mixed effects model with ±1 SE. Acoustic similarity increased over generations, indicating that repetition made the vocalizations easier to imitate with high fidelity. Transcription agreement for the first and last generation imitations of each transmission chain. The mean orthographic distance between the most frequent transcription and all other transcriptions of a given imitation is shown, with error bars denoting ±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, indicating that repeating imitations made them easier to transcribe into English orthography than direct imitations of environmental sounds.

We systematically answered the following questions about the form of the vocalizations and their potential to function as words. First, do vocalizations stabilize in form and become more word-like as they are imitated? 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.

Imitations of environmental sounds became more stable over the course of being repeatedly imitated as revealed by increasing acoustic similarity along individual transmission chains. Research assistants (*N*=5) rated the acoustic similarity of pairs of imitations while blind to all conditions and hypotheses, inter-rater reliability: ICC = 0.76, 95% CI [0.70, 0.81], F(170, 680) = 4.18, *p* < 0.001. Acoustic similarity ratings were fit with a linear mixed effects model predicting similarity from generation with random effects (intercepts and slopes) for each rater and for each seed sound nested within category. Imitations from later generations were rated as sounding more similar to one another than imitations from earlier generations, *b* = 0.10 (SE = 0.03), *t*(11.9) = 3.03, *p* = 0.011 (Fig. 1b). This result suggests that imitations became more stable (i.e., easier to imitate with high fidelity) with each generation.

The possibility that the increase in acoustic similarity was due to transmission chains converging on similar acoustic forms through continuous degradation of the signal was ruled out by additional analyses of acoustic similarity using Mel-frequency cepstral coefficients (MFCCs). These analyses compared acoustic similarity for pairs of sounds sampled from the same and from different transmission chains. We calculated the average acoustic similarity for six types of pairwise comparisons: (1) along transmission chains (consecutive generations), (2) within transmission chains (all pairwise comparisons), (3) all imitations of individual seed sounds, (4) all imitations from the same category of seed sounds, (5) imitations sampled from the same generation between categories, and (6) imitations sampled from consecutive generations between categories. To test whether acoustic similarity was always greater within chains than between them, we fit a linear model predicting acoustic similarity from the type of pairwise comparison coded with Helmert contrasts. The results revealed that acoustic similarity was the highest along consecutive generations of transmission chains (1), followed by comparisons (2) through (6), all *p*'s < 0.01, with imitations from different categories being the least similar to one another (Fig. S2). The ordering of these acoustic similarity measures supports the conclusion that transmission chains were stabilizing on unique acoustic forms.

As an additional test of stabilization, we had English-speaking participants transcribe a sample of first and last generation imitations into English orthography (Fig. S3), and found that last generation imitations were transcribed more consistently than first generation imitations. We collected a total of 2163 transcriptions---approximately 20 transcriptions per sound. Some examples of the transcriptions are presented in Table 1. To measure the similarity among transcriptions, we calculated the orthographic distance between the most frequent transcription and all other transcriptions of a given imitation. The orthographic distance measure was a ratio based on longest contiguous matching subsequences between pairs of transcriptions. We then fit a hierarchical linear model predicting orthographic distance from the generation of the imitation (First generation, Last generation) with random effects (intercepts and slopes) for seed sound nested within category. The results showed 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*(3.0) = -3.62, *p* = 0.035 (Fig. 1c). The same result is reached through alternative measures of orthographic distance for each imitation, such as the percentage of exact string matches, *b* = 0.10 (SE = 0.03), *t*(90.0) = 2.84, *p* = 0.006, and the length of longest substring match, *b* = 0.98 (SE = 0.24), *t*(15.1) = 4.14, *p* < 0.001 (Fig. S4).

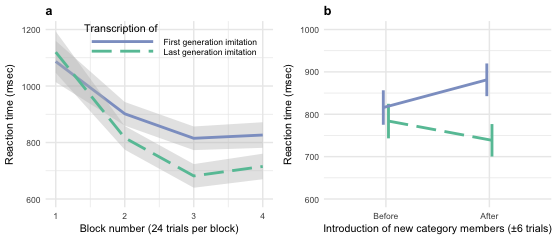
Examples of invented words.

|  |  |  |  |
| --- | --- | --- | --- |
| 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 |

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 is more closely yoked to a particular environmental sound. To the extent that repeating imitations abstracts away the idiosyncrasies of a particular category member25,26, it may also be easier to generalize to new category members. We tested these predictions using a category learning task in which participants learned novel labels as catogry labels of the seed environmental sounds. The novel labels were transcriptions generated either from first or last generation imitations gathered in our experiment.

Participants began by learning through trial-and-error to associate four labels with four categories of environmental sounds. The small number of categories made this an easy task (mean accuracy after the first block of X trials was Y; Fig. S5). 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 and did not differ 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. 2a). 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 process as category labels. We predict that given a harder task (i.e., more than four categories and 16 exemplars) would yield differences in accuracy as well.

Next, we examined whether transcriptions from last generation imitations were easier to *generalize* to novel category exemplars (e.g., are participants who learned that “XXXX” means water-splashing better at generalizing it to another water-splashing sound than participants who learned that “XXXX” means water-splashing. 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. 2b). This result suggests that transcriptions from later generation imitations were easier to generalize to new category members.



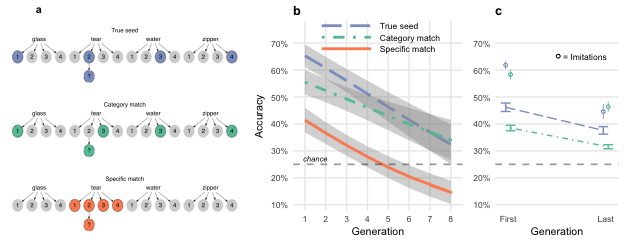
Participants learned novel labels (transcriptions of first or last generation imitations) for categories of environmental sounds through trial-and-error. Mean RTs for correct responses in the category learning experiment with ±1 SE. Participants achieved faster RTs in matching transcribed labels to environmental sounds for labels transcribed from later compared to earlier generation imitations. Cost of generalizing to new category members with ±1 SE. After each block of trials, new environmental sounds were introduced, requiring participants to generalize the previously learned category labels to new category members. There was a generalization cost for first generation labels, but not last generation labels.

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. 3a). All 365 imitations were tested in the three question types depicted in Fig. 3a. 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. 3b. 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 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 imitation for these more difficult questions[[1]](#footnote-1). 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. 3a); 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. 3c). 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.



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. 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 thetrue seed advantage'' (True seed v. Category match) decreased (see main text), suggesting that imitations lose within-category information more rapidly than between-category information. 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.

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.

Imitative (or "iconic") words are found across the spoken languages of the world13,14,19. 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 environment23 and more abstract meanings20, 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.

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 cues25–27.

Even 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 small28,29, 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 languages30. Even English, a language that has been characterized as relatively limited in iconic vocabulary31, is documented to have hundreds of clearly imitative words including words for human and animal vocalizations as well as various types of environmental sounds24,32. 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 states30,33,34. As with onomatopoeia, ideophones are often recognized by naïve speakers as bearing a degree of resemblance to their meaning35.

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 domains20. 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), 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 (Fig. S1). The four final categories were: water, glass, tear, zipper. The final 16 seed sounds can be downloaded from [sapir.psych.wisc.edu/telephone/seeds/all-seeds.zip](http://sapir.psych.wisc.edu/telephone/seeds/all-seeds.zip).

**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. Full instructions are provided in the Supporting Information. Each participant listened to and imitated 4 sounds. Participants were assigned one sound from each of the four categories of sounds drawn at random such that participants were unlikely to imitate 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. For example, recordings with loud sounds in the background were removed, and recordings were trimmed to the length of the imitation prior to the next generation. 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 five research assistants listen to pairs (approx. 300) 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.

To obtain algorithmic measures of acoustic similarity, we used the acoustic distance functions included in Phonological Corpus Tools 36. We computed MFCC similarities between pairs of sounds using 12 coefficients in order to obtain speaker-independent estimates. The relationship between acoustic similarity ratings and algorithmic measures of acoustic similarity is discussed in the Supporting Information.

**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.

Transcriptions were generated from the first and the last three generations of all imitations collected in the transmission chain experiment; that is, not all imitations were transcribed (Fig. S3). Participants also provided transcriptions of the original environmental seed sounds (Fig. S6). 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.

Our primary measure of transcription difference was provided by the SequenceMatcher functions in the difflib package of the python standard library which implements a version of Ratcliff and Obershelp's "gestalt pattern matching" algorithm.

**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 so as 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. S6). 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. Results of matching transcriptions directly of environmental sounds are provided in Fig. S6.

**Data availability.** The data that support the findings of this study, along with all methods, materials, and analysis scripts, are available in public repositories described on the Open Science Framework page for this research at [osf.io/3navm](https://osf.io/3navm).

**Analyses.** Statistical analyses were conducted in R using linear mixed-effects models provided by the lme4 package37. Degrees of freedom and corresponding significance tests for linear mixed-effects models were estimated using the Satterthwaite approximation38. Random effects (intercepts and slopes) for subjects and for items were included wherever appropriate.

# References

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

2. Locke, J. An essay concerning human understanding. in *Readings in the history of psychology* (ed. Dennis, W.) (1948).

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

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

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

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

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

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

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

10. Hockett, C. F. In search of Jove’s brow. *American speech* **53,** 243–313 (1978).

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

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

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

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

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

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

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

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

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

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

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

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

23. Lemaitre, G., Houix, O., Voisin, F., Misdariis, N. & Susini, P. Vocal Imitations of Non-Vocal Sounds. *PloS one* **11,** e0168167–28 (2016).

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

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

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

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

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

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

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

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

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

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

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

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

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

37. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **67,** 1–48 (2015).

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

1. 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. [↑](#footnote-ref-1)