The emergence of words from vocal imitations

Edmiston et al. 10.1073/pnas.XXXXXXXXXX

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. S1), and repeated the norming process again with 4 sounds in each category (Fig. S2). 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.

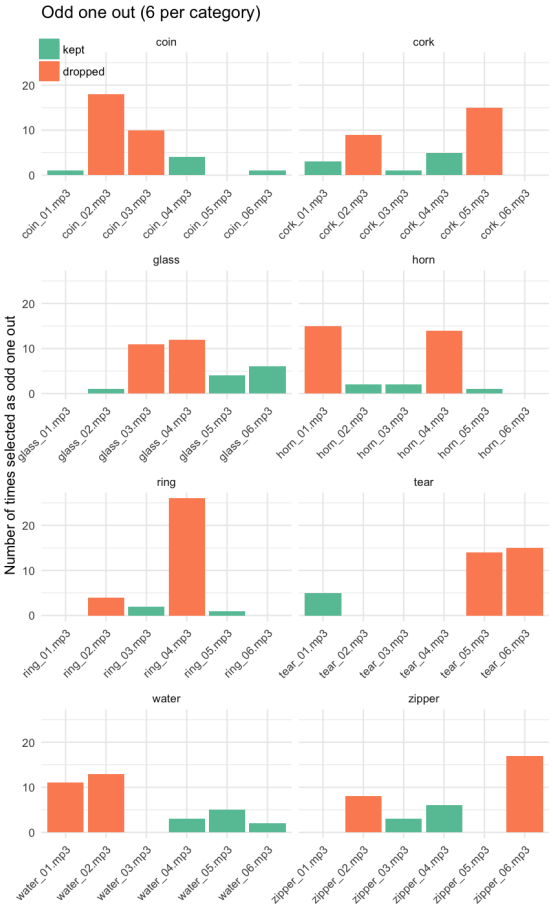

Fig. S1. Results of the first round of seed norming. After collecting these data, two sounds were removed from each category and the norming procedure was conducted again.

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. S3). 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.

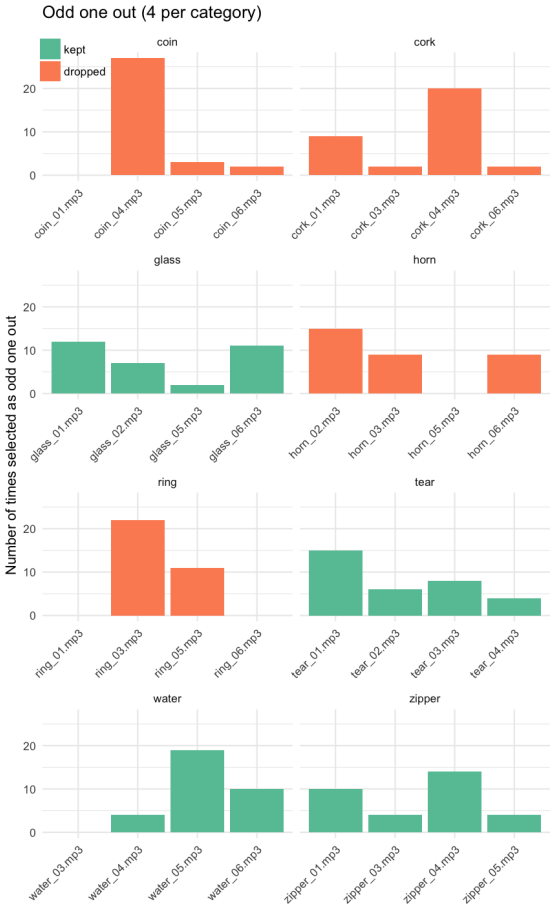

Fig. S2. Results of the second round of seed norming. After collecting these data, four categories of sounds were selected to use in the main experiment.

Fig. S3. 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.

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. [. . .]

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. S4), 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 algorithmic acoustic distance.

Acoustic similarity judgments. Five research assistants rated the similarity between 324 different pairs of imitations. These

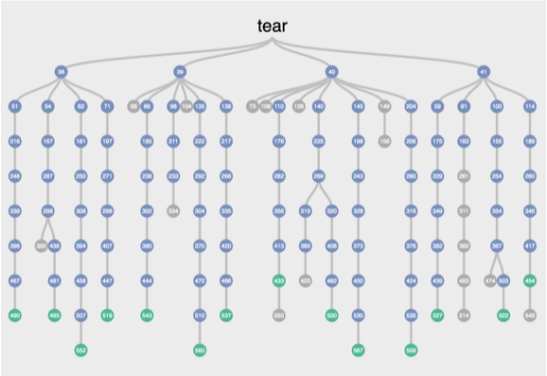

Fig. S4. The interface for monitoring incoming imitations. All imitations were listened to by an experimenter and trimmed to remove extraneous noise. Imitations eligible for the next generation appear in green. Bad quality imitations were rejected (in gray).

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 (36). 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. S5A). 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. S5B). 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. S5C), 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 as the exact wording of the instructions is shown in Fig. S7.

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. S8.

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. S9A).

As can be seen in Fig. S9A, 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. S9B, we show the results of our orthographic distance metric separately for imitations with and without any agreement.

Fig. S9C 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.

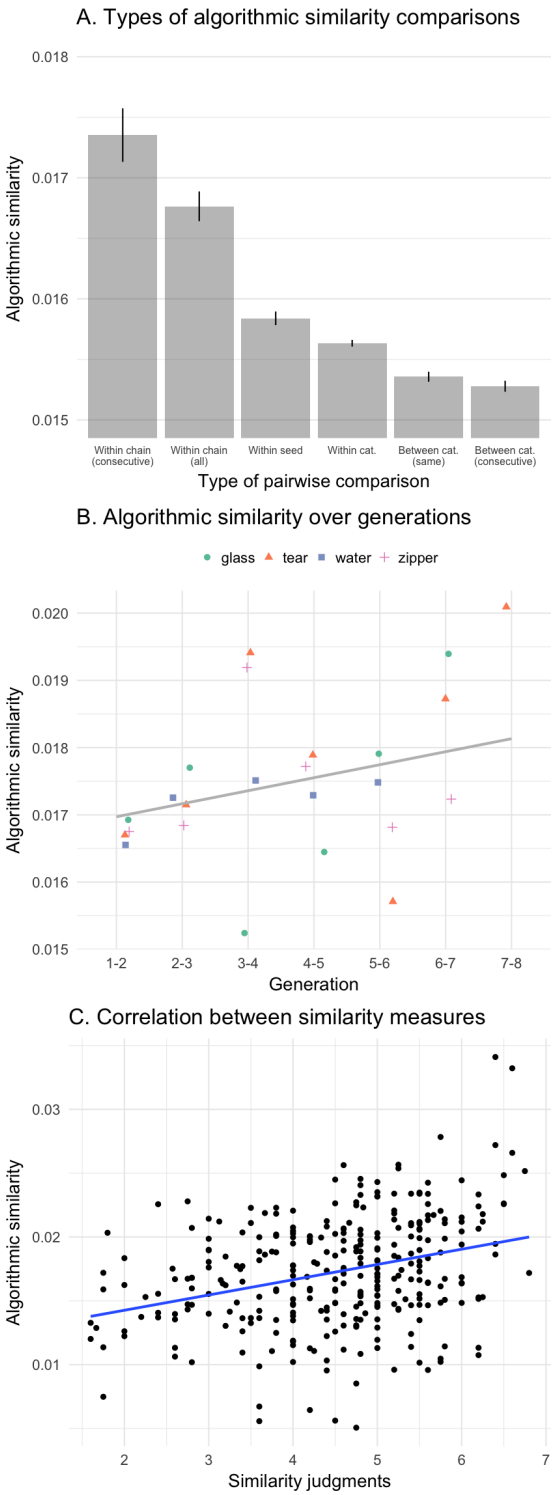

Fig. S5. Algorithmic measures of acoustic distance. A. Average acoustic distance between pairs of sounds grouped by type of comparison. B. Change in algorithmic acoustic distance over generations of imitations. C. Correlation between similarity measures and algorithmic measures.

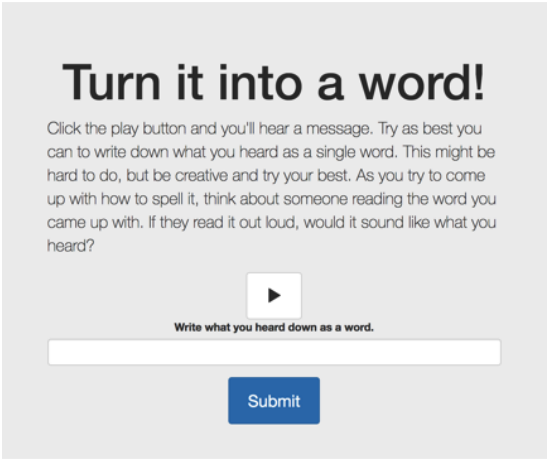

Fig. S6. Interface for collecting transcriptions. Participants listened to an imitation and were instructed to create novel words corresponding to the sound they heard.

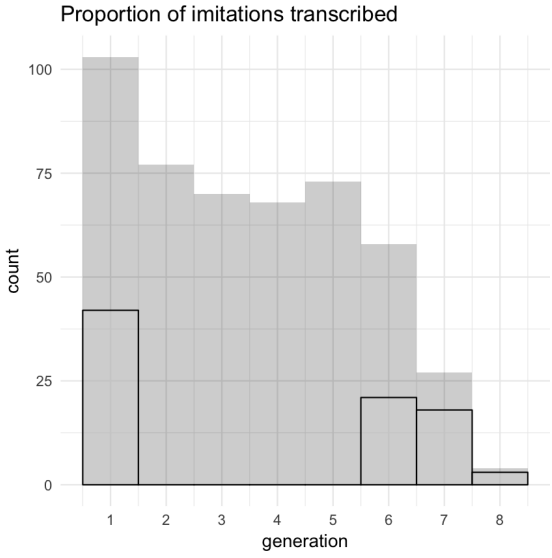

Fig. S7. Proportion of imitations that were transcribed. Gray region indicates the number of imitations collected at each generation. Outlined regions denote the number of imitations that were transcribed. First generation imitations and the last three generations of imitations were transcribed.

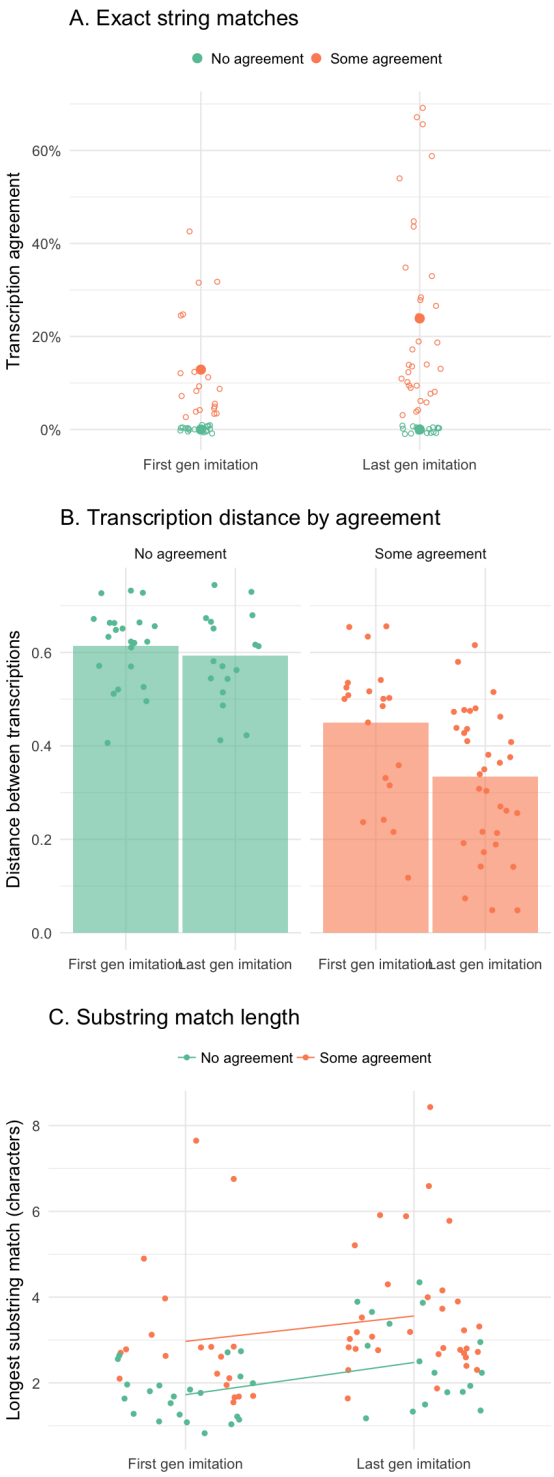

Fig. S8. Alternative measures of orthographic distance. A. Percentage of exact string matches per imitation. B. Orthographic distance separated by whether there was any agreement among the transcriptions of a given imitation. C. Change in the average length of the substring match.

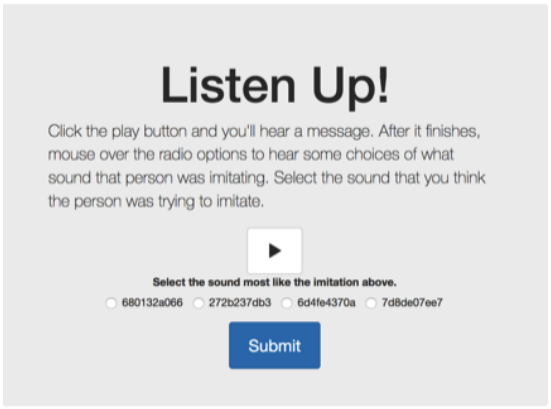

Fig. S9. Interface for matching imitations back to original seed sounds.

Fig. S10. Interface for matching transcriptions back to original seed sounds.

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. S10) or a transcription of an imitation (Fig. S11) 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

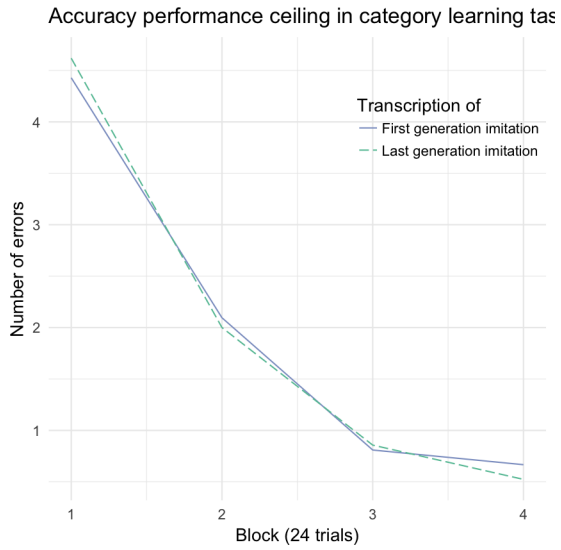

Fig. S11. Mean number of errors by block of 24 trials showing that accuracy performance quickly reached ceiling after the first block of trials.

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. S13A), but the most frequent of them were the easiest to match back to the original seeds (Fig. S13C). When learning these transcriptions as category labels, participants were the fastest to learn them in the first block (Fig. S13B), but they did not generalize to new category members as fast as transcriptions taken from last generation imitations.