Creating words from iterated vocal imitation

Edmiston et al. 10.1073/pnas.XXXXXXXXXX

1. Open data and materials

We are committed to making this research open and reproducible. The R code used to generate the main manuscript as well as all analyses reported in this Supporting Information document is available on GitHub at github.com/lupyanlab/creating-words. All analyses and figures were created in R. The data are available in an R package, which can be downloaded and installed with the following 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 word learning experiment is available at github.com/lupyanlab/learning-sound-names.

2. Selecting seed sounds

Our goal in selecting sounds to serve as seeds in 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 necessary in order to generate 4AFC questions with both between-category and within-category distractors with the appropriate level of counterbalancing across all conditions.

We selected nonverbal environmental sounds because they were less likely to have a lexicalized version already in English and were presumed to be less familiar and more difficult to imitate. In selecting sounds to imitate, we did not want the imitations to be confounded by additional knowledge about the event, e.g., imitating a particular breed of dog's characteristic bark instead of copying the particular sound that was presented as carefully as possible.

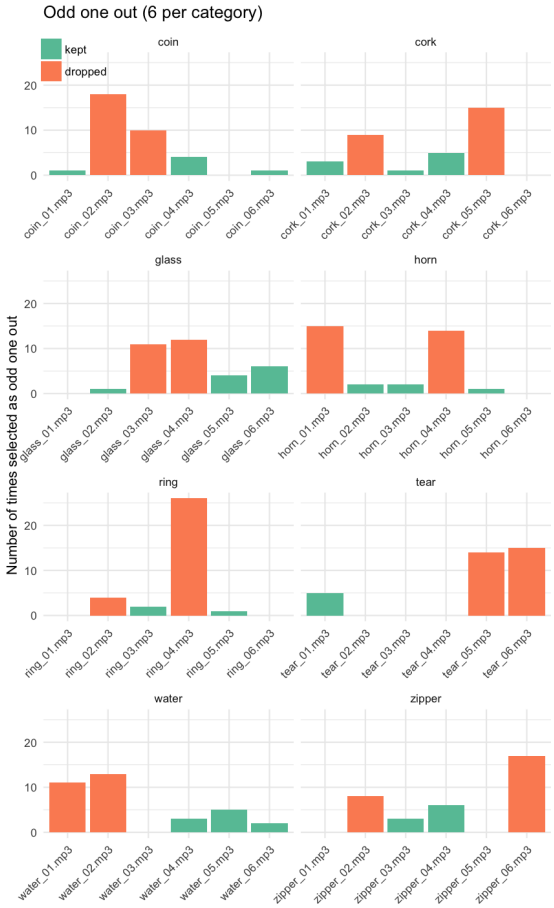


Fig. S1. Results of the first round of seed norming. After collecting these data, two sounds were removed from each category and the norming procedure was conducted again.

Participants in the odd one out norming procedure listened to the 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 sounds in each category that were the most different from the others, and repeated the norming process again with 5 sounds in each category. The resulting sounds that were selected in each category are considered to be a set of equally distinguishable category members.

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

3. Collecting vocal imitations

Participants played a version of the children's game of Telephone via an online interface (Fig. S3). Initially the only action open to players is to hear 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

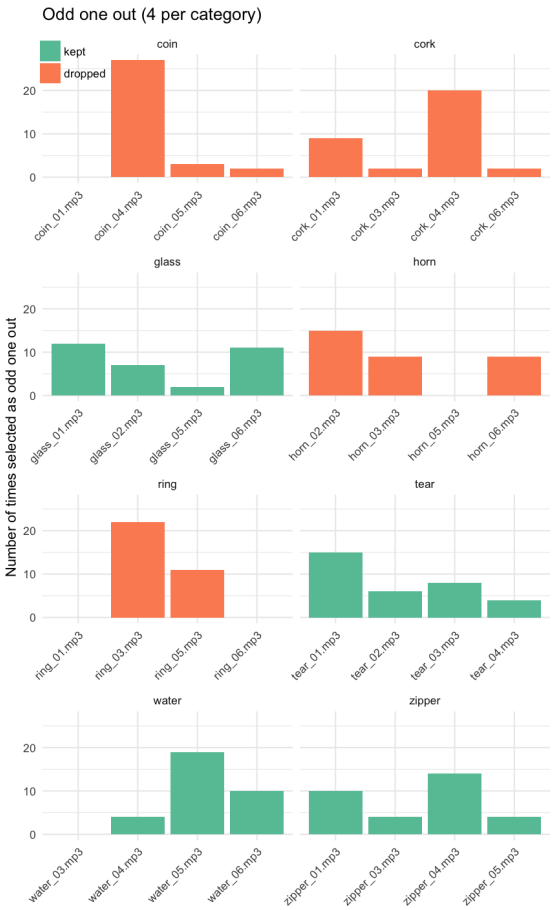


Fig. S2. Results of the second round of seed norming. After collecting these data, four categories of sounds were selected to use in the main experiment.

Table S1. Environmental sounds used as "seed messages".

Category	Exemplar
glass	glass_01.mp3
glass	glass_06.mp3
glass	glass_02.mp3
glass	glass_05.mp3
tear	tear_01.mp3
tear	tear_03.mp3
tear	tear_02.mp3
tear	tear_04.mp3
water	water_05.mp3
water	water_06.mp3
water	water_04.mp3
water	water_03.mp3
zipper	zipper_04.mp3
zipper	zipper_01.mp3
zipper	zipper_03.mp3
zipper	zipper_05.mp3

Fig. S3. The interface for the telephone game. 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 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.

Monitoring incoming imitations. Since the experiment was conducted online, there was a high likelihood 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 needed a way to monitor the imitations as they came in to prevent downstream corruption of transmission chains, and so designed an interactive data visualization (Fig. S4) to check the recordings and exclude ones where necessary. The monitoring also 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, recordings were listened to, trimmed, and in some cases rejected, preventing downstream data corruption. Due to the somewhat random nature of the rejections, all transmissions chains did not reach the expected 8th generation at the same time.

Measuring acoustic similarity. After collecting the imitations in the transmission chain design, the imitations were submitted to an analysis 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. Both measures of acoustic similarity are documented online at github.com/lupyanlab/acoustic-similarity.

Acoustic similarity judgments. Five research assistants rated the similarity between 324 different pairs of imitations. These imitation pairs were sampled from contiguous imitations in the transmission chain design. For example, every message was compared to it's 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

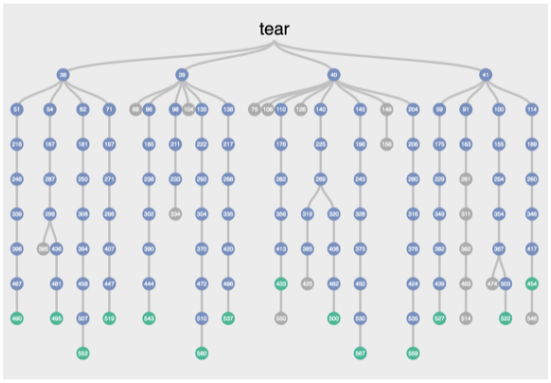
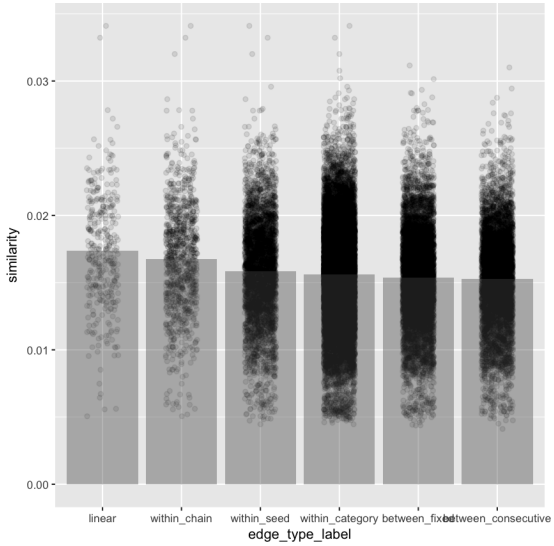


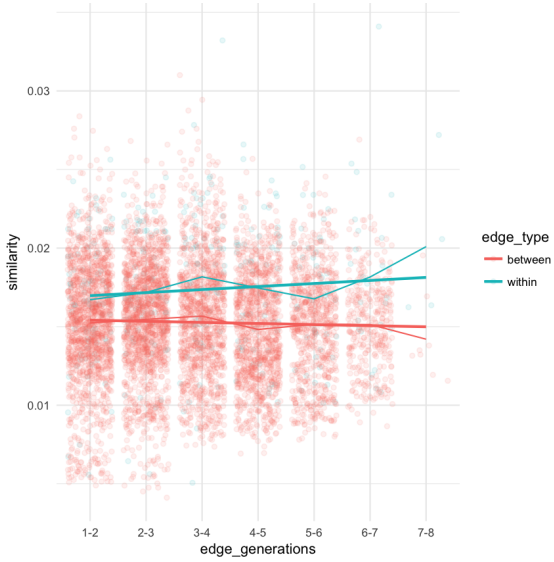
Fig. S4. The interface for monitoring incoming imitations. All imitations were listened to by an experimenter and trimmed to remove extraneous noise. Imitations eligible for the next generation appear in green. Bad quality imitations were rejected (in gray)..

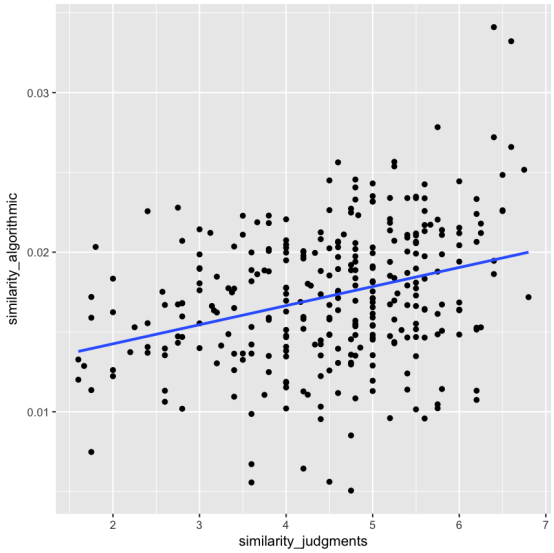
were also blind to the overall generation of the imitations by randomizing generation from trial to trial. To facilitate consistency in rating, pairs of sounds were blocked by category. E.g., participants rated all tearing sounds before moving on to other categories of sounds. The actual instructions 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. If you need to hear the sounds again, you can press 'r' to repeat the trial. If one of the sounds is a non-verbal sound (like someone tapping on the mic), or if you only hear a single sound, or if you are otherwise unable to judge the similarity between the sounds, press the 'e' key to report the error. Pressing 'q' will quit the experiment. Your progress will be saved and you can continue later. Press the SPACEBAR to begin the experiment.

Algorithimic measures of acoustic distance. To obtain algorithmic measures of acoustic distance, we used the acoustic distance functions included in the Phonological Corpus Tools program (34). Using this program, we computed MFCC similarities between pairs of sounds using 12 coefficients in order to obtain speaker-independent estimates.







4. Matching imitations to seeds

To measure the extent to which imitations resembled their seed sound source, we tasked participants with matching the

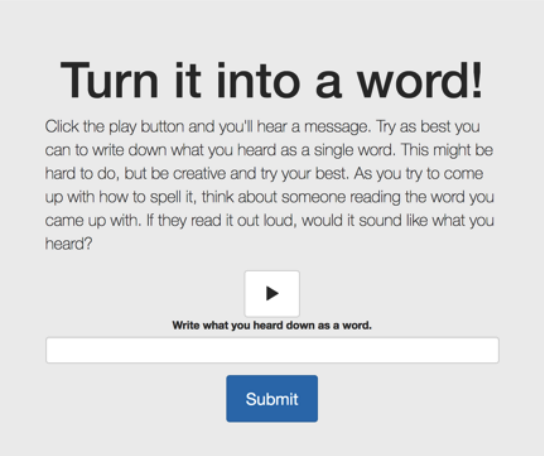
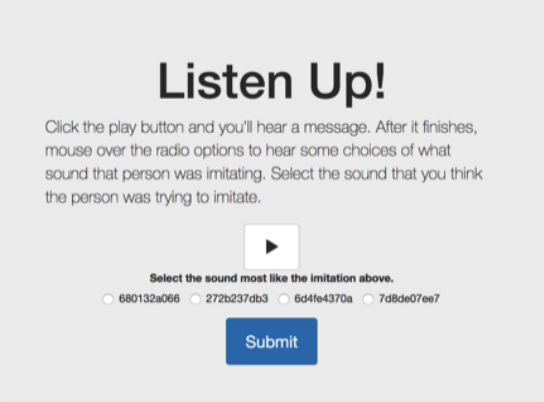


Fig. S5. Interface for collecting transcriptions. Participants were instructed to create novel English words corresponding to the sound they heard.

imitation to it's source relative to other seed sounds used in the experiment. Participants were assigned 4 seed sounds 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. 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.



5. Collecting transcriptions of imitations

For transcriptions, participants were instructed to turn the sound they heard into a word that, when read, would sound much like the imitation.

Fig. S6. Interface for matching transcriptions back to original seed sounds.

6. Matching transcriptions to seeds

In this experiment, rather than matching imitations back to seed sounds, participants read a word formed from a transcription of an imitation back to the seed sound. They were instructed the that word was "invented" to correspond to one of the sounds in their options. As before, participants were assigned 4 seed sounds between-subject to use throughout their experimental session.

7. 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. Then we excluded transcriptions that had less than two unique characters or were over 10 characters long, and sampled from both first and last generation imitations to reach a final set that controlled for overall matching accuracy.

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 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.