The Potential for Iconicity in Vocalization

Marcus Perlman

Max Planck Institute for Psycholinguistics

Gary Lupyan

University of Wisconsin-Madison

Corresponding author:

Marcus Perlman

marcus.perlman@mpi.nl

Abstract

The innovation of iconic gestures is essential to establishing the symbolic vocabularies of signed languages, but what is the potential for iconicity in vocalization and the origins of spoken words? Can people create novel vocalizations that are comprehensible to a naïve listener, without prior convention? We launched a contest in which participants submitted a set of non-linguistic vocalizations for 30 meanings spanning actions, humans, animals, inanimate objects, properties, quantifiers and demonstratives. The winner – who received a monetary prize – was judged by the ability of naïve listeners to successfully infer the meanings of the created vocalizations. Among our participants were eight teams and individuals affiliated with prominent linguistics and language evolution programs in the US and Europe. We report the results from the contest, along with a series of experiments and analyses designed to assess how comprehensible the vocalizations are to naïve listeners, as well as their iconicity and learnability as category labels. Our findings provide a compelling case of the significant potential to use iconic vocalizations to communicate about a wide range of meanings, thereby demonstrating the iconic potential of speech and its origin.

Keywords: vocalization; language evolution; iconicity; sound symbolism

Introduction

In the parlor game charades, players are challenged to shed their spoken language and communicate using gestures. To succeed in creating a signal that successfully communicates the intended meaning to their teammates, players typically make use of *iconicity* – a resemblance between the form of a signal and its meaning. Iconicity enables a receiver to gain some understanding of a signal without relying on a previously learned conventional form-meaning association. In this way, iconicity can ground a new signal and give it meaning.

Outside of the parlor, people face a similar challenge when communicating with someone who speaks a different language, a situation in which iconic gestures can likewise serve to help understanding [1]. With extended interactions, iconic gestures – along with deictic gestures like pointing – can support the formation of more fully-fledged gestural symbol systems. For example, deaf children with hearing parents use iconic gestures as the basis for more symbolic homesign systems [2,3]. And within predominantly deaf communities that originally lack a common language, iconicity plays a crucial role in the emergence of full signed languages [4–6].

While the importance of iconicity to the birth of signed languages is clear, its role in spoken languages is much less so, in large part because spoken languages are so ancient. In this study, we examined the possibility that the words of spoken languages may have been formed in a parallel way to the way many signs were originally created—through a process rooted in iconicity, but in the vocal rather than visual modality. To do this, we tested the extent to which people are capable of creating non-linguistic vocalizations that are effective at communicating various meanings to naïve listeners. We launched a contest—*The Vocal Iconicity Challenge*—in which participants were challenged to communicate a set of basic meanings by inventing novel vocalizations. We assessed the winner of the contest by the ability of naïve listeners to successfully infer the meanings of the created vocalizations.

Background

Iconicity in speech and vocalization

In contrast to the clear influence of iconicity in gesture and sign, many have argued that speech affords a very limited potential for iconicity [1,4,7,8]. For instance, Tomasello (2008) observed that it is difficult to imagine people inventing "vocalizations to refer the attention or imagination of others to the world in meaningful ways – beyond perhaps a few vocalizations tied to emotional situations and/or a few instances of vocal mimicry" [1]. Reaching a similar conclusion from a different perspective, Pinker and Jackendoff (2005: p. 209) argued that vocal iconicity could not be an important factor in the origin of spoken words because "Most humans lack the ability (found in some birds) to convincingly reproduce environmental sounds." They proposed that the human capacity for vocal imitation is essentially limited to "a capacity to learn to produce speech." Within actual spoken languages, Saussure's [9] notion of the arbitrariness of the sign is commonly adopted as essential to the nature of spoken words [10], and the number of iconic and imitative words has been assessed as "vanishingly small" [11]. According to Hockett [8], this is the inevitable consequence of the limited dimensions of speech to afford iconicity.

Some scholars of language evolution have pointed to observations like these – claiming limited potential for iconicity in vocalization and speech – as evidence that the first languages were gestural. On this idea, the first languages originated from iconic gestures that served, eventually, to scaffold arbitrary vocalizations [1,3,4,12,13]. Similar rationale supports an argument for a multimodal division of labor in which gestures and speech co-evolved, but with gesture carrying the iconic load and bootstrapping arbitrary speech [14,15].

However, an improved understanding of the vocabularies of non European languages and increased empirical scrutiny of the “arbitrariness of the sign” dogma have revealed that iconicity in spoken languages is much more pronounced than previously suspected [16–19]. For example, many spoken languages have thousands of ideophones, a distinctly iconic class of words used to express meanings across diverse domains like animate and inanimate sounds, manner of motion, size, visual patterns, textures, inner feelings and cognitive states [17,20,21]. Linguists have also identified many iconic words outside of the ideophone lexical class. For example, across many languages, words expressing *smallness* and related concepts are expressed with high front vowels and *large* concepts with low back vowels [22,23]. This pattern may help explain the differences between typically feminine and masculine personal names [24], and it may also motivate the forms of the indexical words used to refer to proximal and distal referents, such as the translational equivalents of English “here” and “there” [25] and “this” and “that” [25,26]. Words for proximal referents tend to contain front vowels, whereas distal words tend to contain back vowels. Iconicity is also prevalent in some anatomical vocabulary, as languages show a prevalence of nasal consonants in words for “nose” and bilabial consonants for "lip" [27] and recent large-scale analyses of basic vocabulary across thousands of languages have confirmed that some of these relationships between forms and meanings hold across languages [28].

In addition to iconicity in the phonetics of words, a more dynamic form of iconicity in spoken language is found in the intonation, tempo, and loudness – the prosody – of speech. Bolinger [29] suggested that a fundamental function of prosody, especially intonation, is the iconic expression of emotion. Ohala [30] and others [31] have noted that speakers also use prosody to express qualities related to size, dominance, and strength. More broadly, experimental evidence indicates that prosody can enhance the iconicity of ideophones spanning meanings across the senses [32].

Production experiments have also shown that speakers sometimes produce iconic modulations in their prosody when communicating about a range of meanings. For example, speakers have been shown to increase or decrease their tempo when respectively describing a fast or slow-moving event, and to raise or lower their pitch when describing upward or downward movement or when referring to something small or large [33,34]. Iconic prosody may be especially evident in speech directed towards young children. Three adults were asked to produce novel words in infant directed speech, with the meaning paired to one of 12 antonymic adjectives [35]. Analysis of the utterances showed certain consistent differences between the prosodic properties associated with particular meanings, including properties like fundamental frequency, amplitude, and duration. For instance, *strong* was expressed with higher amplitude than *weak*; *happy* with higher pitch, higher amplitude, and a shorter duration than *sad*; and *tall* with a longer duration than *short*. Moreover, naïve listeners were better than chance at selecting a picture matching the original meaning of the word from two alternatives [34].

Inventing novel vocalizations

While the human ability to vocally imitate is often assumed to be poor [10], empirical results paint a different picture. Lemaitre and Rocchesso [36] asked participants to imitate various mechanical and synthesized sounds or to provide verbal descriptions of them. When these were played back to listeners, participants were better at identifying many of the original sounds from the vocal imitation than from the verbal description. A subsequent study found that people are effective at communicating with vocal imitations because they focus on a few salient features of the source, rather than producing a high fidelity representation [37].

In addition to the direct imitation of sounds, recent experiments have shown that people are able to spontaneously invent iconic vocalizations to represent various other kinds of meanings [18]. Participants played a charades-type game in which they took turns improvising non-linguistic vocalizations to communicate meanings from 30 different pairs of antonyms, including words like *alive*, *dead*, *dull*, *sharp*, *hard*, *soft*, *fast*, *slow*, *bad*, *good*, *bright*, and *dark*. Their vocalizations were highly consistent in the particular acoustic features that were used to distinguish contrasting words in more than two thirds of the antonymic pairs. For example, *rough* compared to *smooth* was expressed with aperiodic sounds marked by a lower harmonics-to-noise ratio, *small* compared to *large* with quiet, high-pitched sounds, and *fast* compared to *slow* with loud, high-pitched, quickly repeated sounds.

Other studies have shown that these invented vocalizations are, to some degree, understandable to naïve listeners. One experiment compared the use of non-linguistic vocalization and gesture to communicate 18 items that included emotions (e.g. *disgust*, *tired*), actions (e.g. *throwing*, *chasing*) and objects [38,39]. Compared to the chance rate (5.6%), in the initial block, accuracy was highest for emotions (~60%), next for actions (~40%), and lowest for objects (~10%). In another study, participants took turns for ten rounds producing non-linguistic vocalizations to communicate a set of meanings from nine antonymic pairs of words, including items like *bad*, *good*, *big*, *small*, *down*, *up*, *far*, *near*, *fast*, *slow*, *few*, *many*, *rough*, and *smooth* [40]. With few exceptions, each meaning was expressed with characteristic acoustic properties that distinguished it from each other meaning. In subsequent playback experiments, naïve listeners were better than chance at guessing all but one of the 18 meanings, and for 15 of them, their accuracy was at least 20% and as high 73%, compared to a chance rate of 10%.

Current Study

The work reviewed above shows that (1) iconicity pervades spoken languages much more than previously realized and (2) that people have some ability to invent vocalizations that can be understood by naïve listeners. But just how good are people at doing this? What is the extent of the human potential to ground a symbol system through vocalizations, without the use of gesture? To find out, we conducted the *Vocal Iconicity Challenge!* – a contest in which participants were tasked with creating the most iconic vocalizations for 30 meanings spanning actions, humans, animals, inanimate objects, properties, quantifiers and demonstratives. The iconicity of the vocalizations was evaluated by the ability of naïve listeners to guess their meanings from a set of alternatives. The winning team or individual whose vocalizations were guessed most accurately received a prize of $1000. Submissions included participants affiliated with several prominent linguistics and language evolution programs at universities in the United States and Europe.

Results

Here we report the results from the contest, along with a series of experiments and analyses designed to evaluate the vocalizations for 1) how comprehensible they are to naïve listeners; 2) how iconic they are; and 3) whether iconicity helps naïve listeners learn these vocalizations as category labels. The overarching goal of our analyses was to assess the extent to which people might be able to use iconic vocalizations to ground a spoken symbol system.

First, we analyzed the comprehensibility of the vocalizations by asking naïve listeners to guess their meaning from a set of alternatives. The vocalizations were presented to listeners in two testing conditions (see Methods). In within-category testing, listeners selected the meanings from alternatives in the same broad semantic category (actions, properties, nouns), and in between-category testing, the meanings were selected from choices that included some thematically related alternatives (e.g. *rock*, *pound*, *dull*). We calculated the guessing accuracy for each submitted set of vocalizations. Additionally, we examined guessing accuracy for the different meanings and semantic categories, as well as the errors that guessers made.

Comprehensibility provides one index of iconicity, but we also wanted to assess the degree to which listeners actually perceived a resemblance between form and meaning. Therefore, we next asked naïve listeners to directly rate the degree to which the vocalizations sounded like their intended meaning. Additionally, we wanted to examine iconicity from the perspective of creating and articulating the vocalizations. We did this by measuring the consistency between participants in the vocal qualities they used to represent each meaning [40]. We then tested whether the level of agreement in how to produce an iconic vocalization for a given meaning correlated with the guessing accuracy of naïve listeners.

Finally, we examined whether iconicity helps people learn the vocalizations as labels for categories. Based on the iconicity ratings, we used low-, medium-, and high-iconicity vocalizations as stimuli to test whether naïve listeners were better at learning the meanings of more iconic signals. We manipulated the feedback that learners received – full feedback indicating the correct response or accuracy only – to assess whether iconicity might be especially helpful under more challenging learning conditions.

Comprehensibility of Vocalizations

Figure 1 shows accuracy in within- and between-category testing for each of the 11 submissions. In within-category testing, average accuracy over all submissions was 39.0% (SD = 14.1%), significantly higher than chance (10%), b_0 = 1.64, 95% CI = [1.35, 1.94], z = 11.26, p << 0.001. Accuracy ranged from 58.1% for the top submission to 20.0% for the last-place submission, which was still reliably higher than chance, b_0 = 0.62, 95% CI = [0.057, 1.04], z = 2.55, p = 0.011. In between-category testing, average accuracy was 35.9% (SD = 13.9%), significantly higher than chance (10%), b_0 = 1.48, 95% CI = [1.19, 1.76], z = 10.33, p < 0.001. Accuracy ranged from 56.0% for the winning submission to 12.2% for the last-place submission. The last-place submission was not higher than chance, but the next-to-last place submission (21.5%) did reliably exceed chance, b_0 = 0.71, 95% CI = [0.16, 1.14], z = 2.99, p < 0.003. Overall guessing accuracy was reliably higher in within-category testing than between-category testing, b = -0.18, z = -2.15, 95% CI = [-0.34, -0.016], p = 0.032.

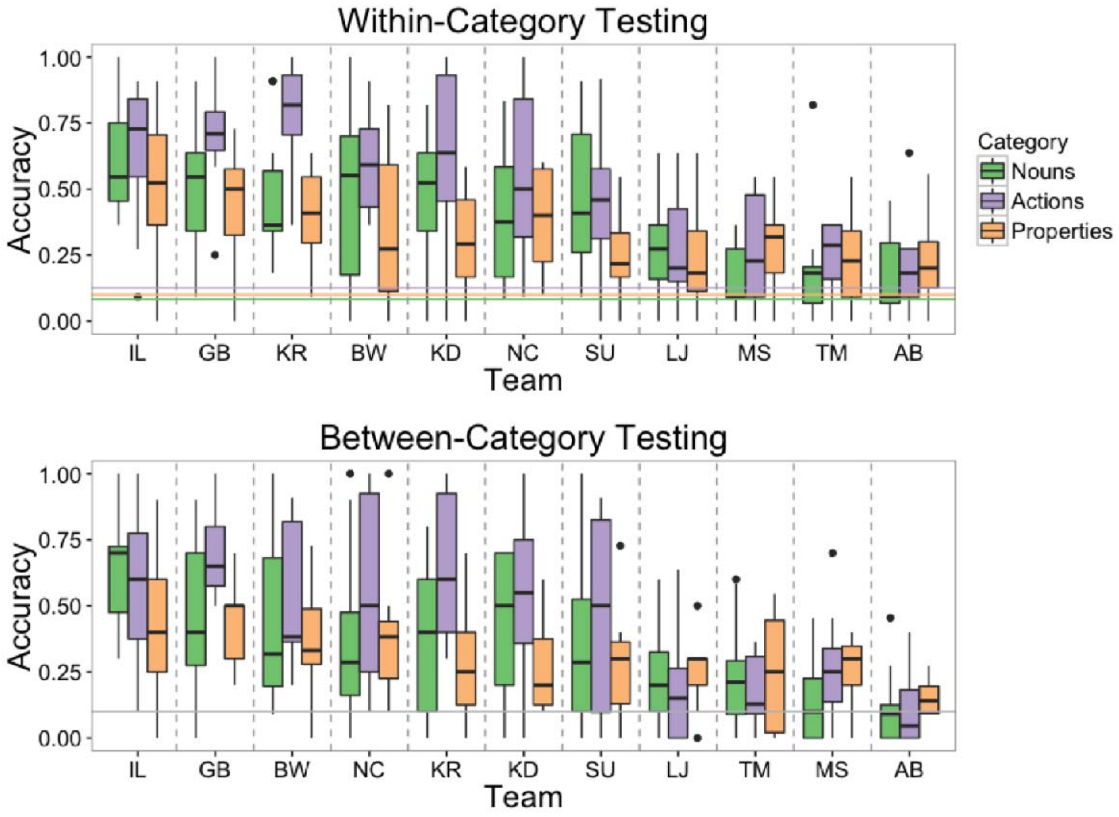


Figure 1. Accuracy for each submission. Top plot shows accuracy in within-category testing. Dashed horizontal lines show chance accuracy rate, with color corresponding to the meaning category. Bottom plot shows between-category accuracy. Black dashed horizontal line indicates chance rate.

Averaged across both testing conditions, accuracy for individual items ranged from 15.4% (SD = 10.2%) for “that” to 72.7% (SD = 33.3%) for “sleep”. Of the 30 meanings, only “that” was not guessed more accurately than chance, b_0 = 0.38, 95% CI =

[-0.18, 0.94], $z = 1.39$, $p = 0.16$. The meaning with the next lowest guessing accuracy, “fruit”, was guessed correctly above chance, $b_0 = 1.01$, 95% CI = [0.1706927 1.2593776], $z = 2.67$, $p < 0.01$ (logistic regression with participant and submission as random effects).

Across both testing phases, the accuracy for actions was 45.6% (SD = 32.0%), for properties 31.8% (SD = 21.5%), and for nouns 36.6%, (SD = 27.9%). Figure 2 shows the results for each meaning in both testing conditions. For each condition, we constructed a logistic mixed effects model to assess the relationship between guessing accuracy and the meaning categories. The model included random intercepts for listener, contestant (i.e., submission ID), and meaning. In within-category testing – in which vocalizations for actions were selected from the 8 alternatives in the set of actions, properties from the 10 properties, and nouns from the 12 nouns – actions were guessed with significantly higher accuracy than properties, $b = -0.80$, 95% CI = [-1.49, -0.11], $z = -2.34$, $p = 0.019$, but with only marginally higher accuracy than nouns, $b = -0.55$, 95% CI = [-1.22, 0.11], $z = -1.69$, $p = 0.09$. In between-category testing, in which all items were selected from 10 alternatives, the meanings of actions were guessed with only marginally higher accuracy than properties, $b = -0.60$, 95% CI = [-1.30, 0.10], $z = -1.73$, $p = 0.08$, and there was not a statistical difference between actions and nouns, $z = -1.24$, p = 0.21.

We next examined whether vocalizations for particular meanings were more likely to be confused with some meanings rather than others. Figure 3 shows confusion matrices for the two phases of testing, with the items ordered roughly according to semantic similarity based on Google’s *word2vec* semantic vectors [41]. The warm-colored diagonals from upper left to bottom right show that listeners most frequently selected the intended meaning of the vocalizations. However, the matrices reveal that some meanings were often confused. For instance, in within-category testing (Figure 3 A-C), vocalizations for “woman” were often confused with “child”, “that” was confused with “dull”, and “many” was confused with “bad”. The between-category matrices (Figure 3 D-F) show a tendency for participants to confuse thematically related meanings, such as “knife” with “cut” and “child” with “small”.

To determine whether listeners were more likely to confuse semantically similar meanings, we used Google’s word2vec semantic vectors as a metric of similarity between pairs of words. These vectors indicate the degree of contextual similarity between two words in a range from 0 (most distant) to 1 (most similar). For each vocalization in each of the two testing conditions, we computed the proportion of trials in which the intended meaning was confused with each possible alternative. For example, in between-category testing, the “gather” vocalization of one submission was confused with “many” in 40% of trials. We then constructed a linear mixed effects model to test whether the semantic similarity between the intended meaning and the confused meaning predicted the proportion of trials in which listeners made this confusion. The model included random intercepts for testing condition, submission, intended meaning, and response. Similarity was a significant predictor of people’s incorrect responses, $\chi_1^2 = 28.0$, $p < 0.001$, $b = 0.07$, 95% CI = [0.05, 0.10], $R^2 = 0.006$ for fixed effects and $R^2 = 0.061$ for fixed and random effects [42]. Similarity was an even stronger predictor of confusability when the proportion of correct responses for each vocalization was included with a similarity of 1, $\chi_1^2 = 2243.8$, $p < 0.001$, $b = 0.33$, 95% CI = [0.31, 0.34], $R^2 = 0.29$ for fixed effects and $R^2 = 0.32$ for fixed and random effects.

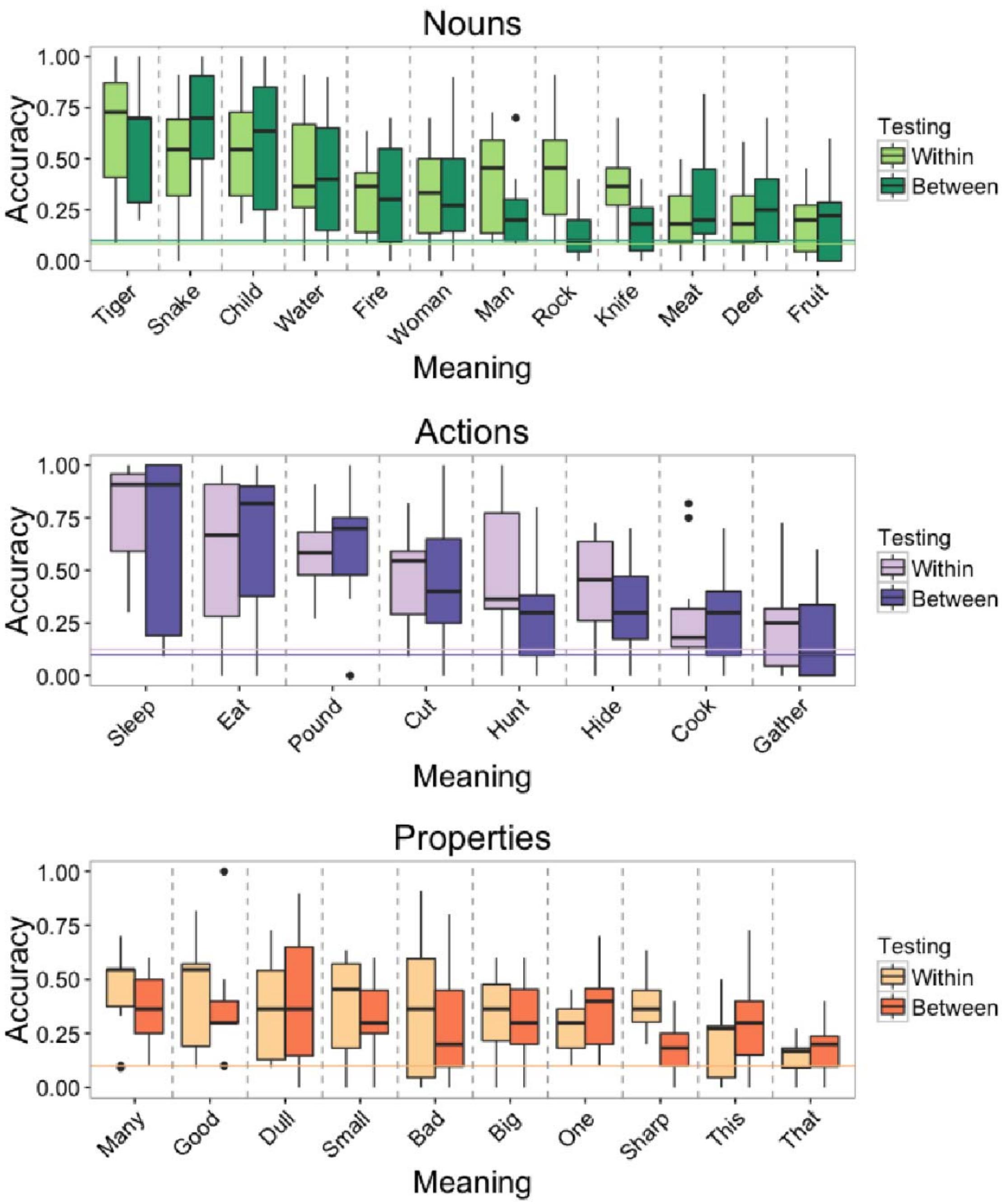


Figure 2. Accuracies for each meaning by category. Top plot shows nouns, the middle plot actions, and the bottom plot properties. Dashed horizontal lines indicate chance accuracy, with color matched to testing round.

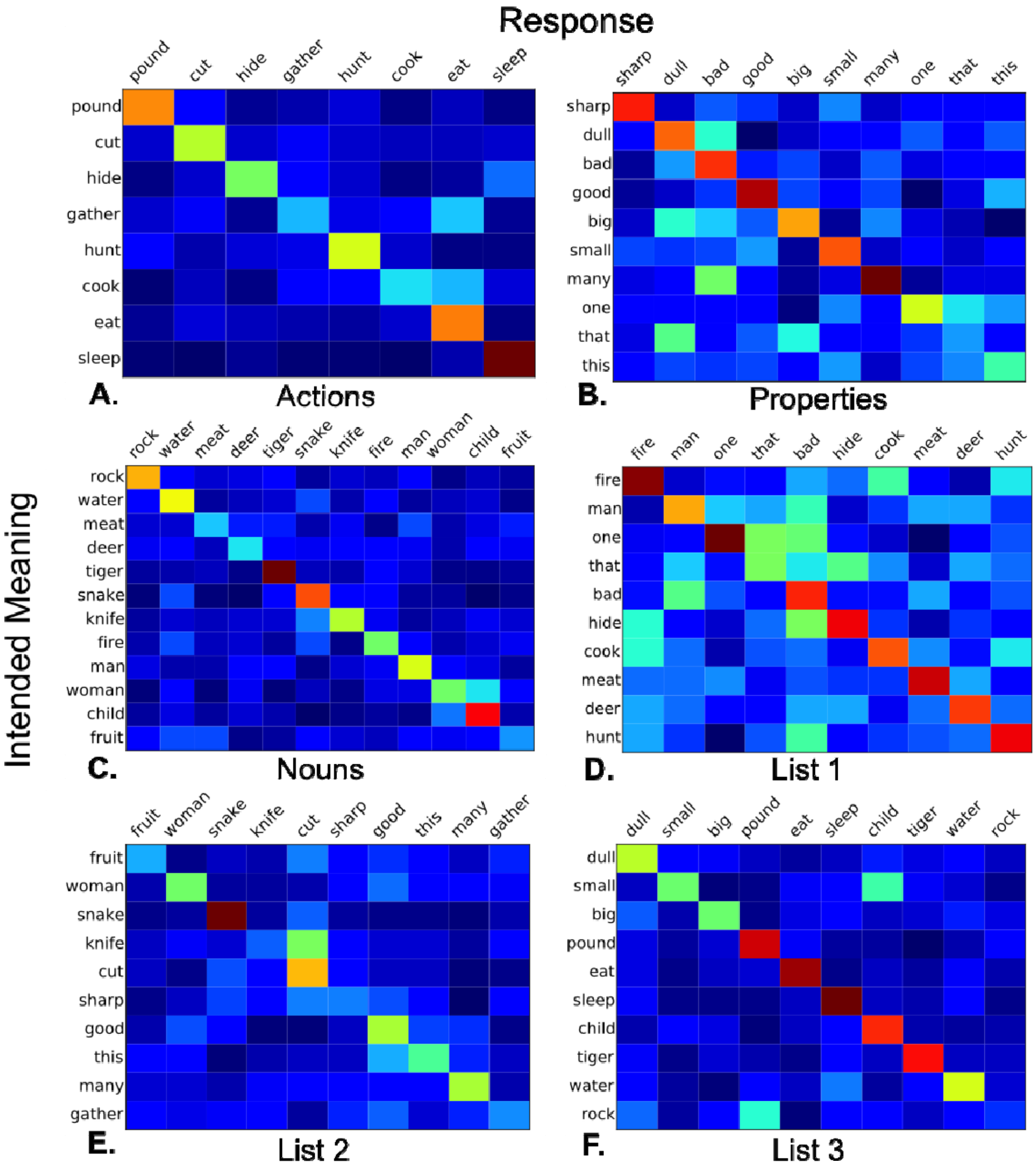


Figure 3. Confusion matrices showing results from the two testing conditions. The y-axis shows the intended meaning of the vocalizations, and the x-axis shows the guessed meanings. The warmer the color of a cell, the more frequently that response was confused for that intended meaning. Items are ordered according to semantic similarity based on Google's word2vec semantic vectors. *A* shows actions, *B* Properties, *C* nouns, *D* list 1, *E* list 2, and *F* list 3. Intended meanings are on the y-axis, and responses are on the x-axis. A-C are ordered in alphabetical order, and D-F are ordered with related meanings next to each other.

Other measures of iconicity

We examined the iconicity of the vocalizations in two additional ways. First, we asked naïve listeners to rate the degree to which the vocalizations "sound like" their intended meaning, thereby providing a more direct evaluation of form-meaning resemblance. Figure 6 shows the distribution of iconicity ratings for each meaning by semantic category. To determine whether the level of iconicity of the vocalizations differed between semantic categories, we constructed a mixed effects model of iconicity rating, with semantic category as a fixed effect, and random intercepts for vocalization, submission, and (rating) subject. Model comparisons showed a marginally reliable effect of semantic category on ratings, $\chi_1^2 = 5.38$, $p = 0.068$. A subsequent model comparing just actions and nouns showed that actions were rated significantly higher in iconicity, $b = -0.46$, 95% CI = [-0.88, -0.04], $p = 0.03$. Next we used a logistic mixed effects model to determine whether the iconicity ratings were a reliable predictor of guessing accuracy. This model included the mean iconicity rating for a vocalization as a main effect, and random intercepts for subject, submission, and vocalization. Figure 4A shows the fit of this model against a scatter plot of accuracy as a function of iconicity rating. Iconicity rating was a reliable predictor of guessing accuracy, $b = 0.55$, 95% CI = [0.48, 0.63], $z = 14.49$, $p < 0.001$; $R^2 = 0.15$ for fixed effects and $R^2 = 0.35$ for fixed and random effects.

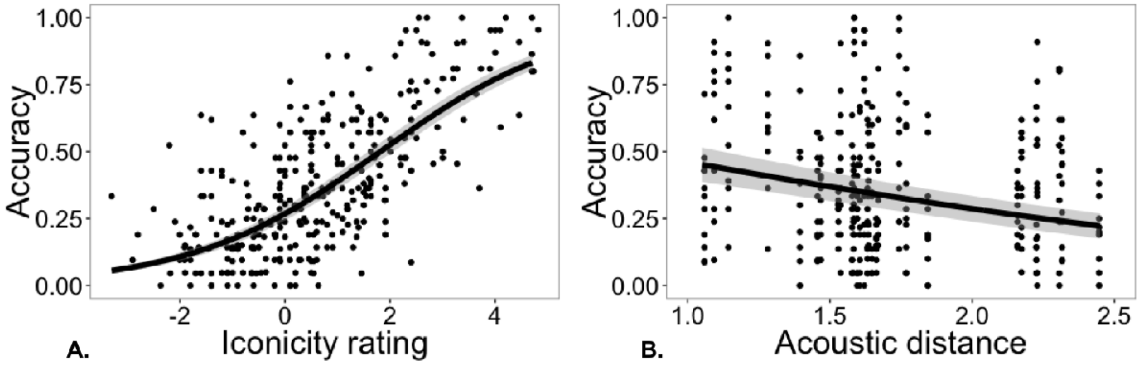


Figure 4. A. Guessing accuracy for each vocalization as a function of its iconicity rating. The line is the fit of a logistic mixed effects model and +/- 1SE of the model estimate. B. Guessing accuracy as a function of the acoustic distance of each meaning. Each point represents a vocalization, but the acoustic distance is for each meaning, resulting in the vertical series of points. The line shows the fit of a logistic mixed effects model, and the grey band indicates the standard error of the model.

We also examined the iconicity of the vocalizations from a production standpoint. For each meaning, we computed the acoustic distance between the vocalizations from each submission using fundamental frequency, duration, intensity, and harmonics-to-noise ratio (see Methods). We then tested whether the degree of similarity of vocalizations (i.e. less acoustic distance) for a given meaning predicted the guessing accuracy of naïve listeners. We reasoned that if producers invented similar-sounding vocalizations for a meaning, then this would indicate an especially strong iconic association, which ought to be reflected in more accurate guessing. To test this, we constructed a logistic mixed effects model, with distance as a main effect, and subject, submission, and vocalization as random intercepts. Figure 4B shows the fit of this model against a scatter plot of guessing accuracy as a function of acoustic distance for each meaning. The analysis showed that acoustic distance was a reliable predictor of guessing accuracy, b = -0.75, 95% CI = [-1.12, -0.39], z = -4.06, p < 0.001. When iconicity rating was added to this model as a main effect, both rating, b = 0.54, 95% CI = [0.46, 0.62], z = 13.97, p < 0.001, and distance, b = -0.35, 95% CI = [-0.64, -0.07], z = -2.41, p = 0.016, were reliable predictors of guessing accuracy. Thus, guessing accuracy was related to both listeners' judgments of form-meaning resemblance – a "sounds like" relationship, and to the level of agreement between the producers of the vocalizations – the degree to which they used similar vocal qualities to express each particular meaning.

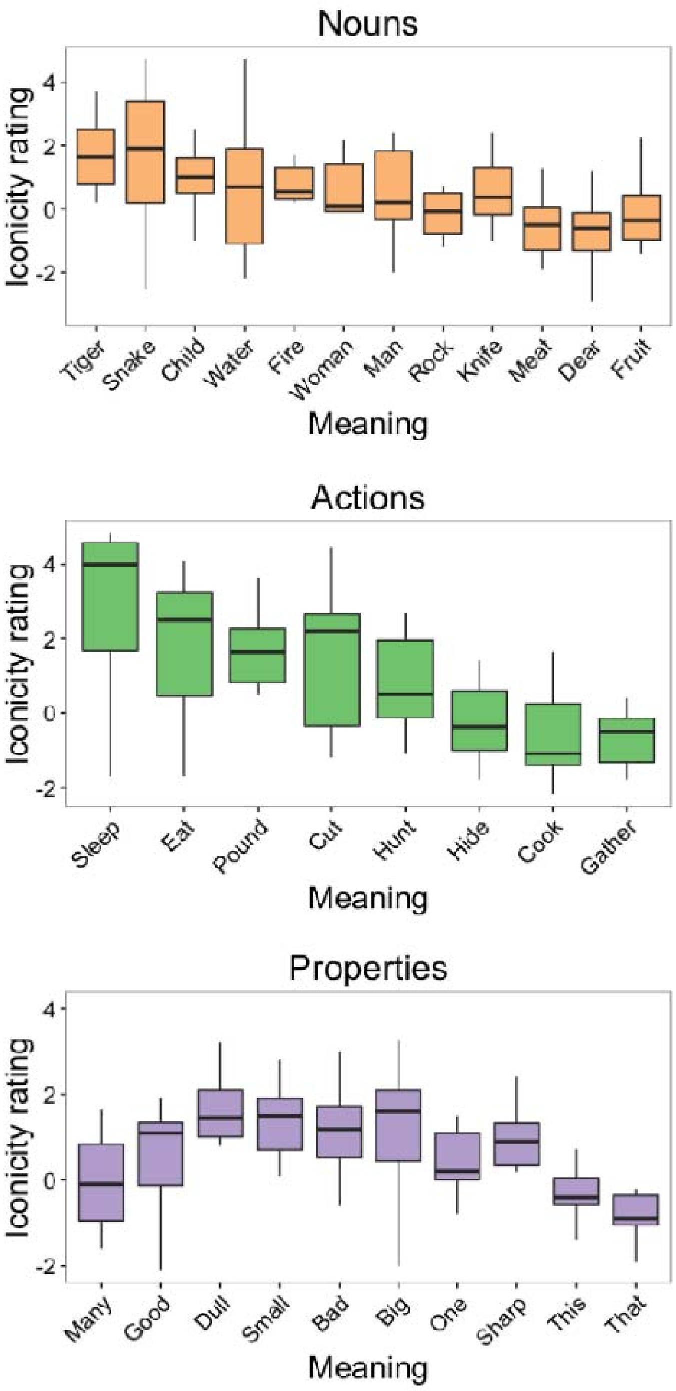


Figure 5. Median, first and third quartiles of iconicity ratings per meaning. Meanings are ordered from lowest to highest guessing accuracy. This ordering shows that iconicity ratings and guessing accuracy were highly correlated, especially for nouns and actions.

Learnability as category labels

Lastly, we conducted a learning experiment to examine whether the iconicity of vocalizations plays a role in the ability of people to learn them as category labels. Participants (University of Wisconsin undergraduates) were tasked with learning to associate twelve vocalizations with 12 noun categories (e.g., fire, man, etc.). They were randomly assigned into a high, medium, or low-iconic group. All three groups completed the same learning task, but learned to associate the categories with vocalizations that were – according to the iconicity ratings we collected – high, medium or low in iconicity. In addition, participants were assigned to one of two feedback conditions: full feedback in which the correct response was indicated, and accuracy only feedback in which their response was only indicated as correct or incorrect.

The results of the learning experiment are shown in Figure 6. To evaluate the results, we constructed a logistic mixed effects model of guessing accuracy. The model included iconicity rating, block, and feedback as main effects, as well as terms for interactions between these variables. Random intercepts were included for subject and item. Not surprisingly, accuracy increased over blocks, showing that that participants were able to learn the categorical meanings of the vocalizations, b = 0.46, 95% CI = [0.43, 0.49], z = 33.70, p < 001. There was a reliable effect of iconicity on accuracy, b = 3.01, 95% CI = [2.15, 3.88], z = 6.80, p < 0.001, such that accuracy was highest in the high iconicity conditions (87.7%), followed by the medium iconicity conditions (58.8%), followed by the low iconicity conditions (47.6%). The model also showed that accuracy was higher in the full feedback conditions (77.8%) compared to accuracy only (51.0%), b = 2.09, 95% CI = [1.74, 2.44], z = 11.71, p < 0.001.

Learning was faster with full feedback than accuracy-only feedback, as supported by a reliable interaction between feedback and block, *b* = 0.28, 95% CI = [0.23,0.34], *z* = 10.74, *p* < 0.001. Learning was faster with higher levels of iconicity, as shown by a significant interaction between iconicity and block, *b* = 0.10, 95% CI = [0.03, 0.17], *z* = 2.95, *p* < 0.01. There was also an interaction between iconicity and feedback, *b* = -1.33, 95% CI = [-2.20, -0.47], *z* = -3.02, *p* < 0.01, suggesting that iconicity provides a larger advantage in learning with accuracy-only feedback by helping the listener to home in on the correct meaning more quickly. The iconicity boost may have been limited in the full-feedback condition as participants reached ceiling performance after only four blocks.

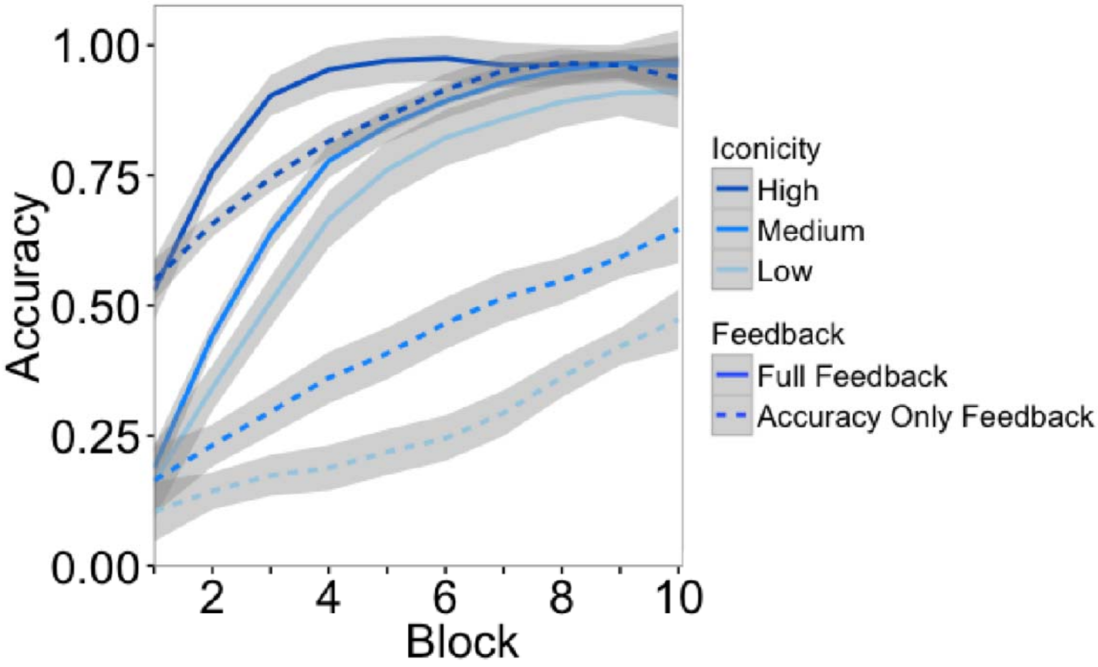


Figure 6. Accuracy over blocks in the learning experiment.

Discussion

Many theories of language evolution have assumed that vocalizations are highly limited in their potential for iconicity – a resemblance between form and meaning, which can help listeners understand the meaning of an unfamiliar signal. Therefore, some theories have reasoned that manual gestures must have played an essential role in bootstrapping the formation of spoken symbols [1,3,4,12,13]. However, evidence from spoken languages [16,19], as well as from experiments [38,40], suggest that people are, in fact, able to invent iconic vocalizations. Our goal in this study was to assess the upper limit of this ability. What is the extent of the human potential to ground a symbol system through vocalizations, without the use of gesture? To investigate this question, we conducted a contest in which participants – incentivized by a monetary prize – competed to devise a set of non-linguistic vocalizations to communicate 30 different meanings to naïve listeners.

After receiving the submissions, we conducted a series of experiments and analyses to evaluate the vocalizations for 1) how comprehensible they are to naïve listeners; 2) how iconic they are; and 3) whether their iconicity helps listeners learn the vocalizations as labels for categories. The findings showed that contestants were able to create vocalizations to successfully communicate basic meanings spanning action verbs, properties (adjectives, quantifiers, demonstratives), and nouns (people, animals, inanimates). For more than half of the submissions, the guessing rate across the 30 vocalizations was greater than 40%, compared to a chance rate of 10%. Over two phases of testing, guessing accuracy was higher than chance for 29 of the 30 meanings. Analysis of the errors guessers made showed that they tended to confuse semantically related words, indicating that a sense of the meaning was often conveyed by the vocalization, even when it did not lead to the correct response.

To further analyze the iconicity of the vocalizations, we asked new listeners to rate them for the degree to which they “sound like” their meaning. These iconicity ratings proved to be a highly significant predictor of guessing accuracy for the vocalizations, suggesting that listeners relied heavily on form-meaning resemblance in making their selections. In addition, we found that when contestants produced more similar sounding vocalizations for a particular meaning, the guessing accuracy for that meaning tended to be higher. Thus, when producers tend to agree on the quality of vocalization to produce for a given meaning, listeners tend to agree on the meaning to interpret from that quality.

Finally, we conducted a learning experiment to examine whether the iconicity of vocalizations plays a role in their learnability as category labels. Are the meanings of more iconic vocalizations easier to learn, especially when informative feedback is limited? The results showed that vocalizations that were higher in iconicity were learned faster than those that were lower in iconicity, particularly when the feedback provided just the accuracy of the response and did not indicate the correct answer. With more iconic vocalizations, listeners were quick to learn their meaning from trial and error after only a few blocks, whereas they often failed to discover the correct meaning of less iconic vocalizations.

Overall guessing accuracy was remarkably high for many of the submissions, and the results suggest that some meanings afford a high potential for iconicity (e.g. *tiger*, *eat*, *many*). However, it is also clear that some meanings afford substantially less. One interesting case is the notably low guessing accuracy of the demonstratives *this* and *that,* which is similar to previous results with the spatial adjectives *near* and *far* [40]. While these meanings may not translate well to iconic vocalizations (but see [26]), they are well suited to pointing gestures. This highlights the potential for iconicity to play complementary roles across modalities: some meanings may better afford iconicity in vocalization and others in gesture.

The most successful submissions – particularly the top six with accuracy rates over 40% – were affiliated with academic programs conducting research in linguistics, psycholinguistics, and language evolution. This raises the possibility that the trained intuitions of language scholars might have been useful for deriving iconic vocalizations that are most understandable to naïve listeners. Another, possibly more important factor, however, is teamwork: four of the top five submissions were created by teams (M = 4.75 participants per team). Both facts point to the possible role of deliberation and interaction in devising successful iconic vocalizations. This is consistent with previous findings with an iterated version of the vocal charades task, which found that participants produced vocalizations that were more understandable to naïve listeners with repeated interactions [40]. Thus, the full expressive potential of vocal iconicity may not be spontaneously available to communicators, but may be sharpened by deliberation and interaction.

One point of qualification of our findings is that the study was limited to English speakers – both in the contestants who produced the vocalizations and in the participants who listened to them. Yet, compared to many psycholinguistic experiments, our contestants were fairly heterogeneous – from across the US, as well as from the UK and Poland, including two submissions from native German speakers and one from native Polish speakers. Is it possible that listeners relied not on iconicity, but on arbitrary conventions shared among the English-speaking population? To a degree, this concern is mitigated by the contest rules, which did not allow spoken emblems or onomatopoeia.

Additionally, the findings that the iconicity ratings were a strong predictor of guessing accuracy and learning suggest that people were tuned in to the resemblance between form and meaning, and that this resemblance played a role in their performance. Nevertheless, future research is required to examine any cultural variability in the patterns we observed here.

Combined, the findings from our contest provide one of the most compelling demonstrations to date of how iconic vocalizations can enable interlocutors to establish understanding through vocalizations in the absence of conventions, thereby demonstrating the iconic potential of speech. Along with other recent studies of iconicity in the production of vocalizations (e.g. Perlman et al., 2015), including vocal imitation [37], our results complement the accumulating evidence of iconicity in the vocabularies [20,28,43], grammar [44], and prosody [33,35] of spoken languages. Taken together, the results support the hypothesis that iconicity in human communication is not limited predominantly to gesturing and signed languages, but also plays an important role in our vocal communication, including speech [16–19]. This newly emerging understanding of iconicity as a widespread property of spoken languages suggests iconicity may also have played an important role in their origin. An intriguing possibility is that many of the now arbitrary words in modern spoken languages may have originated from the innovation of iconic vocalizations.

Methods

Contest

Participants

We recruited contestants by advertising the contest on the Internet, including calls through www.replicatedtypo.com and through an announcement on the Protolang 4 conference website. We received 11 submissions: seven from individuals and four from teams. Nine of the submissions came from the United States, one from the United Kingdom, and one from Poland. There were two submissions by native German speakers, and one by native Polish speakers. Eight submissions came from researchers in relevant academic departments (e.g. Linguistics, Language Evolution, Psychology), and three from individuals not affiliated with academic institutions.

Stimuli

Our set of stimuli consisted of 30 basic meanings: 8 action verbs (*cook*, *cut*, *eat*, *gather, hide*, *hunt*, *pound*, *sleep*), 12 nouns referring to people (*child*, *man*, *woman*), animals (*snake*, *tiger*, *deer*), and inanimate things (*fire*, *fruit*, *knife*, *meat*, *rock*, *water*), and 10 properties including adjectives (*bad*, *big*, *dull*, *good*), quantifiers (*one*, *many*), and demonstratives (*that*, *this*). By design, some of the meanings were thematically related, e.g., *cut*, *knife*, and *sharp*.

Submissions

Contest instructions and other details were made available at the original contest website: http://sapir.psych.wisc.edu/vocal-iconicity-challenge/ and in the Supplementary Materials. Potential contestants were asked to submit a set of recorded vocalizations for each of the 30 meanings together with a brief statement explaining the rationale of each vocalization. Vocalizations were defined as sounds “produced by the vocal apparatus”.

Contestants were permitted to produce imitative sounds, but not allowed to use recognizably onomatopoeic words or conventional emblems, e.g., “booo” (bad), “nyam nyam” (eat), or “roar” (the sound of a tiger). We verified that none of the submissions violated this rule in any clear way. The most marginal case was the use of sibilants for *snake*; however, in each instance, the sound was elaborated beyond any standard convention.

The instructions stipulated that each vocalization should be no longer than two seconds. Twenty-four (7.3%) of the submitted vocalizations exceeded this duration (16 from the last-placed contestant AB). The top contestant followed this restriction, and so it did not interfere with determining the winner of the contest. For completeness, we report results with all of the vocalizations included.

Determining the winner

The winner of the contest was determined by experiments that tested how well naïve listeners could guess the meaning of the vocalizations from each submission. We tested the vocalizations in two independent phases consisting of within-semantic category and between-semantic category testing (see below). As per the instructions, the top five submissions in the first phase of testing advanced to the second phase, and the submission with the highest overall guessing accuracy in the second phase was crowned the winner. In all of the analyses presented here, vocalizations from all 11 submissions were subjected to both testing conditions.

Comprehensibility of vocalizations

Participants

We recruited 708 participants through Amazon Mechanical Turk to serve as listeners: 366 in within-category testing and 342 in between-category testing. We aimed to test each vocalization with 10 participants in each phase, but this number was sometimes inadvertently exceeded. Participants were restricted to be in the USA and were only permitted to participate once in the study.

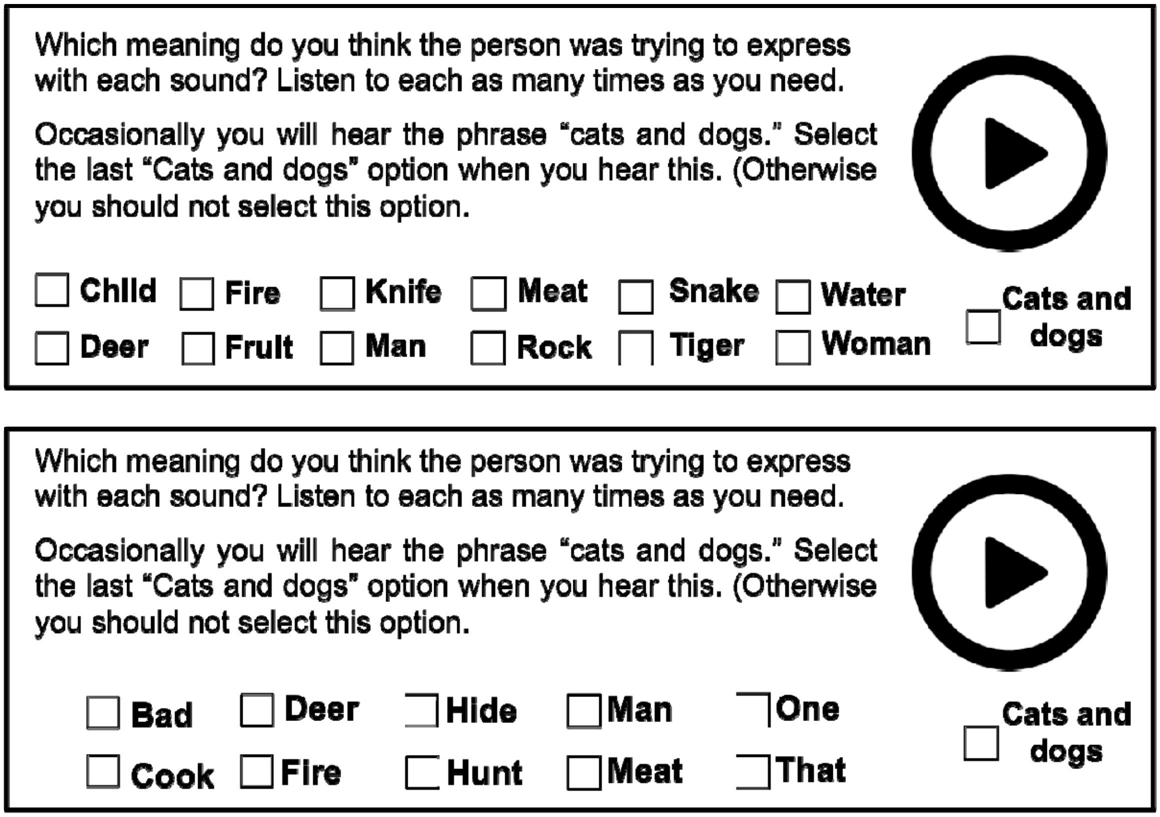
Stimuli

The stimuli were the recorded vocalizations submitted by the contestants in the contest. Eleven contestants each produced vocalizations for 30 meanings, for a total of 330 vocalizations.

Design and procedure

Participants listened to a set of vocalizations from a single contest submission, and selected the meaning of each from a list of alternatives. They could listen to each vocalization as many times as they needed before making their choice. The vocalizations were presented in random order. To ensure that listeners properly attended to the task, we also included catch trials with the spoken phrase “cats and dogs”, along with a corresponding option. Four participants did not respond correctly to these catch trials and were excluded from further analysis.

The vocalizations of each submission were tested separately in two conditions (*within* and *between*). Figure 7 shows trial schematics from each testing condition. In the *within* condition vocalizations were presented with foils from within the same word class. Thus the meaning of a vocalization for an action was selected from 8 alternatives, a vocalization for a property from 10, and a vocalization for noun from 12. On *between* testing trials, the foils came from between all word classes. Meanings were placed into three lists of 10, with some thematically related meanings deliberately included to increase difficulty (e.g. *rock*, *pound*, *dull* and *knife*, *cut*, *sharp*; Figure 3 shows the meanings used in each list).



Which meaning do you think the person was trying to express with each sound? Listen to each as many times as you need.

Occasionally you will hear the phrase "cats and dogs." Select the last "Cats and dogs" option when you hear this. (Otherwise you should not select this option.

□ Child □ Fire □ Knife □ Meat □ Snake □ Water □ Cats and dogs
□ Deer □ Fruit □ Man □ Rock □ Tiger □ Woman

Which meaning do you think the person was trying to express with each sound? Listen to each as many times as you need.

Occasionally you will hear the phrase "cats and dogs." Select the last "Cats and dogs" option when you hear this. (Otherwise you should not select this option.

□ Bad □ Deer □ Hide □ Man □ One □ Cats and dogs
□ Cook □ Fire □ Hunt □ Meat □ That

Figure 7. A schematic of the testing method. The top panel shows a within-category trial with nouns. The bottom panel shows a between-category trial.

Other measures of iconicity

Iconicity ratings

To assess the degree to which listeners perceived a resemblance between the forms of the vocalizations and their intended meanings, we asked naïve listeners to directly rate the degree to which each one "sounds liked what it means".

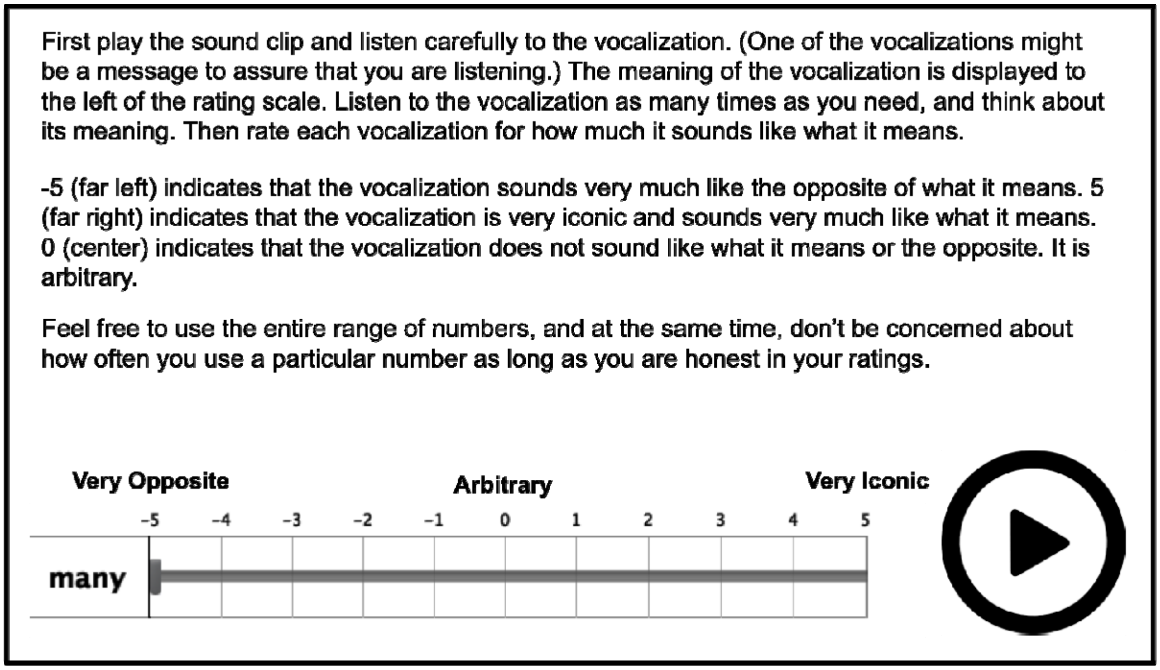
Participants. We recruited 229 new participants through Amazon Mechanical Turk. The target was to acquire 10 ratings for each vocalization (i.e. 10 raters for each of the 22 lists), and this number was exceeded in a few cases. Participants were restricted to be in the USA and were only permitted to participate once in the study.

Stimuli

The stimuli were the 330 recorded vocalizations submitted by the 11 contestants in the contest.

Design and procedure. The vocalizations were separated into 22 pseudo-randomized lists of 15 different meanings. Each list contained vocalizations from each of the 11 contestants, with 4 contestants repeated twice.

Using a procedure similar to [43], we collected ratings of the iconicity of the vocalizations. Participants listened to the vocalizations in a random order, with the meaning of each one displayed. For each vocalization, they were asked to indicate “how much it sounds like what it means” on a scale of -5 (very opposite) to 0 (arbitrary) to 5 (very iconic).



First play the sound clip and listen carefully to the vocalization. (One of the vocalizations might be a message to assure that you are listening.) The meaning of the vocalization is displayed to the left of the rating scale. Listen to the vocalization as many times as you need, and think about its meaning. Then rate each vocalization for how much it sounds like what it means.

-5 (far left) indicates that the vocalization sounds very much like the opposite of what it means. 5 (far right) indicates that the vocalization is very iconic and sounds very much like what it means. 0 (center) indicates that the vocalization does not sound like what it means or the opposite. It is arbitrary.

Feel free to use the entire range of numbers, and at the same time, don't be concerned about how often you use a particular number as long as you are honest in your ratings.

Figure 8. Schematic of the interface used for iconicity ratings.

Similarity per meaning

To measure the degree to which contestants produced similar vocalizations for each meaning, we devised a similarity metric based on the acoustic properties of the vocalizations. We used Praat phonetic analysis software (Boersma, 2001) to measure the duration, pitch, intensity, and harmonics to noise ratio of the vocalizations. The onset and offset of each vocalization was marked by hand, and a script was used to automatically measure the four properties. The formula for the similarity metric is shown below. For each meaning (M), the standard deviation of each property (DUR = duration, PIT = pitch, INT = intensity, HNR = harmonics to noise ratio) was divided by the mean value of that property over all meanings. These four values were then added together, providing a metric of overall similarity of the vocalizations for each meaning.

$$\mathrm{SIM.DUR}_{\mathrm{M}} = \mathrm{SD.DUR}_{\mathrm{M}} \ / \ \mathrm{MEAN.DUR}_{\mathrm{Total}}$$

$$\mathrm{SIM.PIT}_{\mathrm{M}} = \mathrm{SD.PIT}_{\mathrm{M}} \ / \ \mathrm{MEAN.PIT}_{\mathrm{Total}}$$

$$\mathrm{SIM.INT}_{\mathrm{M}} = \mathrm{SD.INT}_{\mathrm{M}} \ / \ \mathrm{MEAN.INT}_{\mathrm{Total}}$$

$$\mathrm{SIM.HNR}_{\mathrm{M}} = \mathrm{SD.HNR}_{\mathrm{M}} \ / \ \mathrm{MEAN.HNR}_{\mathrm{Total}}$$

$$\mathrm{SIM}_{\mathrm{M}} = \mathrm{SIM.DUR}_{\mathrm{M}} + \mathrm{SIM.PIT}_{\mathrm{M}} + \mathrm{SIM.INT}_{\mathrm{M}} + \mathrm{SIM.HNR}_{\mathrm{M}}$$

Learnability of vocalizations as category labels

Participants

We recruited 87 undergraduate students from University of Wisconsin-Madison to participate in exchange for course credit.

Stimuli

The stimuli consisted of a subset of the recorded vocalizations from the contest. Owing to the difficulty of depicting actions and properties with static images, we restricted our materials to the 12 nouns. For each of the 12 nouns, we selected 10 different pictures that clearly depicted its referent (e.g. 10 pictures of a fire for *fire*, 10 pictures of fruit for *fruit*).

For each meaning, three vocalizations were selected on the basis of the iconicity ratings. These included the vocalization with the lowest mean iconicity rating, the one with the median rating, and the one with the highest rating. These comprised the low, medium, and high iconicity conditions, respectively.

Design and procedure

Participants were randomly assigned to one of the three iconicity conditions and, within each condition, one of two feedback conditions as described below. On each trial, participants heard a vocalization and attempted to select its meaning from 12 pictures (arranged in a 3x4 grid) representing each of the different meanings. Participants completed 10 blocks with each block including vocalizations from each of the 12 categories, in a random order. The position of the images and the specific exemplar used for the meaning were randomized with the stimulation that all category exemplars of a target category were used as targets at some point for each participant. Participants selected the image depicting the category of the vocalization made their selection by clicking on one of the pictures with the mouse. They then received feedback on their selection according to one of two randomly assigned conditions. In the *full feedback* condition, the correct response was explicitly indicated by highlighting the correct image. In the *accuracy only* condition, the participant was informed only whether their response was correct or incorrect. Because each image served as a target exactly once throughout the experiment, participants could not respond correctly simply by associating a sound with a specific image.

Statistical analyses

Statistical analyses with mixed effects models were conducted using the lme4 package version 1.1-10 [45] in R version 3.2.3 [46]. Significance tests of continuous outcomes were calculated using $\chi 2$-tests that compared the model likelihoods with and without the factor of interest. Significance tests of binomial outcomes used the z-values associated with the logistic mixed-effect models.

Data availability

Data and analysis scripts, including vocalizations from the contest, are available through the Open Science Framework at osf.io/x9w2z.

Authors' Contributions: M. Perlman and G. Lupyan devised the experiments, analyzed the data, and wrote the manuscript.

Competing Interests: We have no competing interests.

Funding: This work was funded in part by an NSF INSPIRE Award 1344279 to GL.

References

1. Tomasello, M. *Origins of Human Communication*. (The MIT Press, 2008).
2. Goldin-Meadow, S. *The Resilience of Language: What Gesture Creation in Deaf Children Can Tell Us about how All Children Learn Language*. (Psychology Press, 2003).
3. Goldin-Meadow, S. What the hands can tell us about language emergence. *Psychon. Bull. Rev.* 1–6 (2016). doi:10.3758/s13423-016-1074-x
4. Armstrong, D. F. & Wilcox, S. E. *The Gestural Origin of Language*. (Oxford University Press, 2007).
5. Klima, E. & Bellugi, U. *The Signs of Language*. (Harvard University Press, 1979).
6. Meir, I., Padden, C., Aronoff, M. & Sandler, W. Competing iconicities in the structure of languages. *Cogn. Linguist.* **24,** (2013).
7. Corballis, M. C. *From hand to mouth: the origins of language*. (Princeton University Press, 2003).
8. Hockett, C. F. In Search of Jove's Brow. *Am. Speech* **53,** 243–313 (1978).
9. Saussure, F. de. *Course in General Linguistics*. (Open Court Publishing, 1983).
10. Pinker, S. & Bloom, P. Natural language and natural selection. *Behav. Brain Sci.* **13,** 707–727 (1990).
11. Newmeyer, F. J. Iconicity and Generative Grammar. *Language* **68,** 756–796 (1992).
12. Arbib, M. A. *How the Brain Got Language: The Mirror System Hypothesis*. (Oxford University Press, 2012).

13. Levinson, S. C. & Holler, J. The origin of human multi-modal communication. *Phil Trans R Soc B* **369,** 20130302 (2014).
14. Kendon, A. Semiotic diversity in utterance production and the concept of 'language'. *Philos. Trans. R. Soc. B Biol. Sci.* **369,** 20130293–20130293 (2014).
15. McNeill, D. *How language began: gesture and speech in human evolution*. (Cambridge University Press, 2012).
16. Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H. & Monaghan, P. Arbitrariness, Iconicity, and Systematicity in Language. *Trends Cogn. Sci.* **19,** 603–615 (2015).
17. Imai, M. & Kita, S. The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philos. Trans. R. Soc. B Biol. Sci.* **369,** 20130298–20130298 (2014).
18. Perlman, M. & Cain, A. A. Iconicity in vocalization, comparisons with gesture, and implications for theories on the evolution of language. *Gesture* **14,** 320–350 (2014).
19. Perniss, P., Thompson, R. L. & Vigliocco, G. Iconicity as a General Property of Language: Evidence from Spoken and Signed Languages. *Front. Psychol.* **1,** (2010).
20. Dingemanse, M. Advances in the Cross-Linguistic Study of Ideophones: Advances in the Cross-Linguistic Study of Ideophones. *Lang. Linguist. Compass* **6,** 654–672 (2012).
21. Nuckolls, J. B. The case for sound symbolism. *Annu. Rev. Anthropol.* **28,** 225–252 (1999).

22. Haynie, H., Bowern, C. & LaPalombara, H. Sound Symbolism in the Languages of Australia. *PLOS ONE* **9,** e92852 (2014).
23. Ultan, R. Size-sound symbolism. in *Universals of Human Language* (1978).
24. Pitcher, B. J., Mesoudi, A. & McElligott, A. G. Sex-biased sound symbolism in English-language first names. *PloS One* **8,** e64825 (2013).
25. Tanz, C. Sound Symbolism in Words Relating to Proximity and Distance. *Lang. Speech* **14,** 266–276 (1971).
26. Johansson, N. & Zlatev, J. Motivations for Sound Symbolism in Spatial Deixis: A Typological Study of 101 Languages. *Public J. Semiot.* **5,** 3–20 (2013).
27. Urban, M. Conventional sound symbolism in terms for organs of speech: A cross-linguistic study. *Folia Linguist.* **45,** (2011).
28. Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F. & Christiansen, M. H. Sound–meaning association biases evidenced across thousands of languages. *Proc. Natl. Acad. Sci.* **113,** 10818–10823 (2016).
29. Bolinger, D. *Intonation and Its Parts: Melody in Spoken English*. (Stanford University Press, 1986).
30. Ohala, J. J. The frequency code underlies the sound-symbolic use of voice pitch. in *Sound Symbolism* 325–347 (Cambridge University Press, 1994).
31. Pisanski, K. & Bryant, G. A. The evolution of voice perception. in *The Oxford Handbook of Voice Studies* (eds. Eidsheim, N. S. & Meizel, K. L.) (Oxford University Press, in press).

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

33. Perlman, M., Clark, N. & Johansson Falck, M. Iconic Prosody in Story Reading. *Cogn. Sci.* **39,** 1348–1368 (2015).

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

35. Nygaard, L. C., Herold, D. S. & Namy, L. L. The Semantics of Prosody: Acoustic and Perceptual Evidence of Prosodic Correlates to Word Meaning. *Cogn. Sci.* **33,** 127–146 (2009).

36. Lemaitre, G. & Rocchesso, D. On the effectiveness of vocal imitations and verbal descriptions of sounds. *J. Acoust. Soc. Am.* **135,** 862–873 (2014).

37. Lemaitre, G., Houix, O., Voisin, F., Misdariis, N. & Susini, P. Vocal Imitations of Non-Vocal Sounds. *PLOS ONE* **11,** e0168167 (2016).

38. Fay, N., Arbib, M. & Garrod, S. How to Bootstrap a Human Communication System. *Cogn. Sci.* **37,** 1356–1367 (2013).

39. Fay, N., Lister, C. J., Ellison, T. M. & Goldin-Meadow, S. Creating a communication system from scratch: gesture beats vocalization hands down. *Front. Psychol.* **5,** (2014).

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

41. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed Representations of Words and Phrases and their Compositionality. in 3111–3119 (2013).

42. Nakagawa, S. & Schielzeth, H. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods Ecol. Evol.* **4,** 133–142 (2013).

43. Perry, L. K., Perlman, M. & Lupyan, G. Iconicity in English and Spanish and its relation to lexical category and age of acquisition. *PLoS ONE* **10,** e0137147 (2015).

44. Givón. Iconicity, isomorphism, and non-arbitrary coding in syntax. in *Iconicity in syntax* (ed. John Haiman) (John Benjamins Publishing, 1985).

45. Bates, D., Maechler, M., Bolker, B. & Walker, S. lme4: Linear mixed-effects models using Eigen and S4. (2014).

46. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2014).