What do we mean when we say gestures are more expressive than vocalizations? An experimental and simulation study

Šárka Kadavá (kadava@leibniz-zas.de)
Leibniz-Centre General Linguistics, Berlin
Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen
University of Göttingen

Aleksandra Ćwiek (cwiek@leibniz-zas.de)
Leibniz-Centre General Linguistics, Berlin

Susanne Fuchs (fuchs@leibniz-zas.de)
Leibniz-Centre General Linguistics, Berlin

Wim Pouw (wim.pouw@donders.ru.nl)
Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen

Abstract

In this work we discuss the complexity surrounding the origins of human language. We focus on the debate between gesture-first and vocalization-first theories. While some evidence supports the idea that gestures played a primary role in early communication, others argue that vocalizations are equally expressive. We think that methodological differences and biases in the choice of concepts may contribute to the challenge of comparing these modalities directly. For example, to what extent does selecting a certain concept from a semantic category matter to reproduce an effect? This and similar questions are explored in a data-driven way. First, we provide ratings on imagined expressibility of 207 concepts from an online experiment showing that people tend to rate gesture modality as better in expressing meaning compared to vocal modality. Second, we use the Bayesian posterior predictive distribution of these ratings to simulate new experiments where we vary the number of participants, number of concepts, and semantic categories to investigate how robust is the difference between gesture and vocal modality. Our results show that gesture modality is reliably different (i.e., affords higher expressibility) than vocal modality. However, the difference between the two is limited in terms of effect size (medium sizes by common standards) so one may question whether this difference is meaningful for bigger claims about early language evolution. This study further provides valuable information for further research on how to select stimuli and how to set up one's design in a balanced way.

Keywords: Expressibility, Gesture, Vocalization, Language Evolution, Simulation, Effect sizes

Introduction

Language evolution is one of the 'hardest problems in science' (Cooperrider, 2020). Explaining how gestures and/or vocalizations contributed to first meaningful exchanges, and tracing the continuity from our closest living relatives to communication today remains a complex challenge. A significant body of literature assumes that language originated from gestures, referred to as 'gesture-first theories' (Hewes, 1976; Tomasello, 2010). Among a range of arguments, gesture-first views are motivated by case studies showing that some human-reared big apes can learn up to 350 signs (Gardner & Gardner, 1969). This supposedly supports the theory as there is no evidence that great apes are able to acquire and communicate with a comparable repertoire using speech. In addition, research on non-conventionalized signaling (Fay et al., 2022) reported that human adults can better express selected concepts using silent gestures in comparison to novel vocalizations. Finally, children can easily and rapidly (in 30 min) create and communicate novel gesture language during social interaction (Bohn, Kachel, & Tomasello, 2019). Whether the same applies to vocalization is left unknown.

Conversely, some argue that vocalizations are not expressively limited and can convey certain concepts as effectively as gestures (Perlman & Cain, 2014). Human adults from diverse cultural and linguistic backgrounds can comprehend novel vocalizations above chance levels (Ćwiek et al., 2021). Additionally, although non-human animals may possess a limited vocal repertoire compared to humans, they can also devise and learn new vocalizations, such as when perceiving a novel flying object as a potential predator (Wegdell, Hammerschmidt, & Fischer, 2019).

Comparisons favoring one perspective over the other can be somewhat problematic due to methodological differences and the potential biases of research perspectives (Slocombe, Waller, & Liebal, 2011; Liebal, Slocombe, & Waller, 2022; Hoeschele, Wagner, & Mann, 2023). Slocombe et al. (2011) carried out a systematic search of the literature published between 1960–2008 on spontaneous primate communication and reported that gesturing in primates was historically often investigated in social play situations in captivity while vocalizations were frequently recorded in the wild in the context of predator defense. Moreover, most researchers focus on one modality, and only 5 percent of the studies investigated two modalities, making a direct comparison between the gesture and vocal domain almost impossible. In a follow-up review 10 years later (Liebal et al., 2022), the same authors found that not much has changed regarding the focus on unimodal signals. The published studies on multimodal communication in primates even decreased slightly.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Our impression of human communication science is similar in that direct comparisons between novel vocalizations and gestures are rather rare (Fay, Lister, Ellison, & Goldin-Meadow, 2014; Macuch Silva, Holler, Ozyurek, & Roberts, 2020) and may not be without biases. Note that a bias can occur without malintent and can have a variety of sources. One of the possible biases we identify here is the choice and the number of concepts that could be potentially better expressed in one or the other modality. This bias is especially likely to arise because of the time constraints of an experiment, i.e., a handful of concepts might be produced in a given modality.

The literature on novel vocalizations and gestures reminds us of each modality's different affordances for various concepts. Nölle (2021) discusses that gestures are to some degree shaped by environmental factors, such as the visibility of the referent. Fuchs and Ćwiek (2022) argue that visual concepts work best with gestures while auditory concepts, like 'clock', may be better exploited in the auditory domain. More distant communication or communication at night in the dark may require the vocal modality while closer communication may work successfully with gestures. Perlman and Cain (2014) suggest that vocalization could be as good, if not better, in conveying sounds, emotions, and other physiologically correlated acoustics. These semantic categories often relate to sex and size, which can further inform about other social-cultural correlatives such as politeness, confidence, and authority. In a recent cross-linguistic study by Ćwiek et al. (2021), it was shown that some semantic categories are better than others (*actions* > *entities* > *properties* > *demonstratives*). Lievers and Winter (2018) discuss how different parts of speech (i.e., lexical categories) in English relate to certain semantic concepts, e.g., concepts of sounds are frequently expressed as verbs. Finally, Zhou, van der Ham, de Boer, Bogaerts, and Raviv (2024) show that sight and sound exhibit differences in distributed statistical learning – for instance, duration information was better learned in the auditory modality, while visual modality served better for learning spatial information.

All in all, we can assume that the choice of concepts – based, for instance, on their semantic category – may inherently favor expressibility in a certain modality. Importantly, this directly raises a question about the reported effect sizes of experimental studies and the inferences we draw from them. What does it mean when a researcher claims, for instance, that 'gesture is more expressive than vocalization'? Do we take such claims to be true even when there is a very low effect size? A specification is thus needed in what effect size constitutes a meaningful difference. Further, we need to specify how important concepts are in obtaining supposed differences, which also informs about how much room there is for potential bias.

The current work has two aims: First, we provide empirical insights about the imagined expressibility of participants in the vocal, gesture, and multimodal domains for a comprehensive set of 207 concepts. While most of the aforementioned studies investigate expressibility in terms of guessing accuracy, we operationalize it here as a self-reported imagined expressibility of meaning in a certain modality. This allows us to increase the amount of data we can collect in terms of sample size and amount/diversity of concepts. These expressibility ratings may be a valuable resource for the community when selecting concepts for an experiment.

The second aim of our work is to use these gesture and vocal expressibility ratings to simulate new experiments and test the robustness of the primacy of one modality over the other. In other words, we want to quantify the magnitude of the difference between the two modalities. Additionally, we investigate potential biases that could occur due to limited resources in recruitment, time, or simply due to selection criteria applied when picking stimuli.

Part I: Obtaining and Analyzing the Expressibility Ratings

Methods

Material We used a list of concepts initially developed for a cross-cultural study, encompassing 200 items. This list included 100 items from the Leipzig-Jakarta list (Tadmor, 2009) and an additional list of 100 items. We chose the Leipzig-Jakarta list because it contains concepts found in almost all human cultures. The authors selected concepts that were least likely borrowed from other languages and had a preference for older than younger concepts. The additional list was added based on various criteria such as sensory ratings, iconicity, abstractness, and valence ratings of English words, spanning a wide range of semantic categories and parts of speech (Winter, Lupyan, Perry, Dingemanse, & Perlman, 2023). For our purposes, we separated concatenated concepts found in the Leipzig-Jakarta list (e.g., 'he/she/it', 'leg/foot') into individual concepts to avoid colexification issues (François, 2022), resulting in a total amount of 207 concepts. All concepts were then translated from English into German by a native speaker and checked for clarity by another native speaker.

Online experiment In an online experiment built with PsychoPy2 (Peirce et al., 2019) and deployed by Pavlovia.org, we engaged native German speakers to assess their imagined ability to express a concept without using language. They were instructed to rate on a continuous scale, how well they think they could express the meaning of a concept using: (a) only vocalizations, (b) only gestures, or (c) a combination of both. The scale was banded and only two extreme points were indicated, ranging from 'very bad' to 'very good', to prevent biases and allow for high precision (Matejka, Glueck, Grossman, & Fitzmaurice, 2016). Each participant evaluated 20 randomly selected concepts for each modality, resulting in 60 ratings in total. They were recruited through Clickworker.de and received monetary compensation.

Dataset and statistical analyses All statistical analyses were carried out in `R` (R Core Team, 2023) using the following libraries `tidyverse` (Wickham & Wickham, 2017), `broom`

(Robinson, 2014), `brms` (Bürkner, 2017), `cmdstanr` (Gabry, 2021), `HDI Interval` (Dezeure, Bühlmann, Meier, & Meinshausen, 2015), `tidybayes` (Kay, 2020), `loo` (Vehtari, Gelman, Gabry, & Yao, 2021) and `BayesFactor` (Morey, Rouder, Jamil, & Morey, 2023).

We excluded ratings that were more than 3 standard deviations from the mean response time (0.6% of the data) and removed data from one participant who kept responding with almost identical rating values. After these exclusions, the participant pool consisted of 248 individuals (1 diverse, 3 not specified, 141 males, 103 females; average age 40, range 18–70) with 12,854 data points, and between 18–24 ratings per concept on average.

The ratings on a scale from -1 to 1 were transformed to a 0 to 1 scale for modeling purposes. For the analysis, we employed four Bayesian hierarchical zero-one inflated beta models. Each model, consisting of four chains, was run for a total of 8,000 samples post-warmup. All models converged, indicated by $\hat{R}$ values of 1.00. We selected the best-fit model from the four models using leave-one-out cross-validation. This model considered expressibility ratings as a function of modality (on three levels, with the combined modality as the baseline). It incorporated by-concept intercepts and slopes for modality and applied the same structure for the *phi*, *zoi*, and *coi* components.[1]

Results

The model results, displayed in Table 1, indicate that the concepts generally received the highest ratings in combined (i.e., multimodal) condition. Ratings in both gesture and vocal modality are reliably lower.[2]

Table 1: Statistical output of the Bayesian hierarchical zero-one inflated beta model: Intercept = multimodal.

parameter	estimate	s.e.	CrI	$p(\beta < 0)$
Intercept	0.35	0.3	(0.29, 0.42)	1.00
gesture	–0.11	0.03	(–0.17,–0.04)	1.00
vocal	–0.81	0.04	(–0.89,–0.73)	1.00

The modeling script and complete list of 207 concepts, including translations, are available at the OSF repository.

Additionally, we sampled posterior predictive distributions for each concept per modality, using `rstanarm` R-package (Goodrich, Gabry, Ali, & Brilleman, 2024). The resulting matrix was used to simulate the experiments outlined in the next section.

[1] *zoi* defines the proportion of data that belong to a modeled cluster, *coi* defines the proportion of data that belong to a modeled noise, and *phi* defines the precision of distinguishing between modeled cluster and noise.

[2] Note that we report on all three modalities despite excluding multimodality from our further analysis. This is mainly to stay consistent with our modeling procedure.

Part II: Simulating the Effect of Concept Choice

Motivation

Natural experiments assessing expressibility require elaborate designs as they need to include productions and perceptions of gestural/vocal utterances. They are therefore naturally limited by the number of participants that can be recruited, and critically, are naturally limited by the number of possible concepts that can be incorporated into the experiment. These resource limitations prevent assessing a comprehensive number of diverse concepts that would allow for conclusions about a supposed expressive superiority of one modality over another or an entire conceptual space. Most experiments incorporate about 16 to 36 concepts per participant (using both within- and between-subject designs). Since both the number of participants and the number of stimuli are assumed to be sources of random variation (Westfall, Kenny, & Judd, 2014), they directly relate to the statistical power of the subsequent analysis.

The expressibility ratings in Part I provide us with an opportunity to simulate different experiments with different sample sizes and different concept selections, to quantify the likelihood of whether a difference between gesture and vocal communication is observed, and more importantly, to quantify the likely magnitude of the effect. Importantly, we can also assess whether selecting concepts from particular categories that are popular in previous research may have inadvertently biased results as compared to experiments that include concepts from another set of categories. As seen in Figure 1, categories differ in terms of the distribution overlap between the gesture and vocal modality. While simulations are certainly not the final arbiter on these questions, they do provide information that can guide our expectations over and above lab experimentation.

Methods

Based on the reasoning sketched above, the following parameters are varied in our simulations: participants (N_P = 10, 15, 20), number of concepts (N_c = 12, 18, 24), and category choice (no, determined, random).

Simulations In $Simulation_1$, we do not constrain the selection of concepts by any categorical criterion. A certain number of concepts C is randomly selected from the list. For each concept, the number of data points representing the number of participants P, is drawn from the posterior predictive distribution matrix for the vocal and gesture modality.

In $Simulation_2$, the selection of concepts is limited to the three categories used previously, i.e., *action*, *object*, and *emotion*. The number of concepts C is randomly selected from each of the categories.[3] For each concept, the number of data

[3] For the simulations with 24 concepts, i.e., 8 concepts per category, we did not have enough items in the category *emotion*. For that purpose, we created a quasi-item 'x' to which we assigned two – gesture and vocal – distributions with a mean and standard deviation of expressibility in the respective modality within the category.

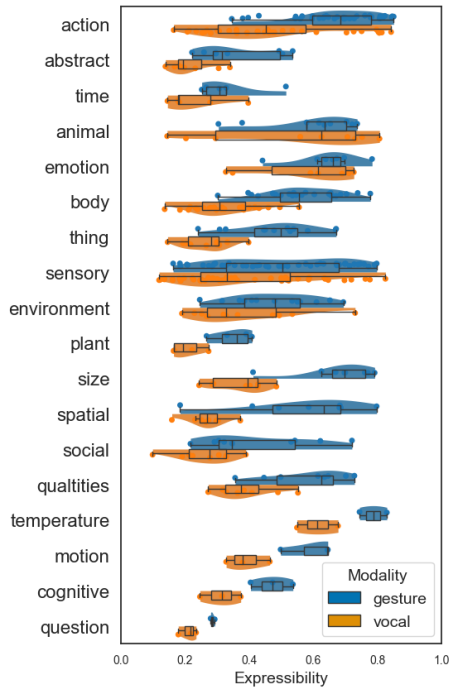

Figure 1: Distributions of gesture and vocal expressibility, stratified by semantic categories.

points P is drawn from the posterior predictive distribution matrix for both vocal and gesture modality.

In $Simulation_3$, the selection of concepts is limited by semantic categories, but the categories themselves are picked randomly. The categories are the following: *sensory*, *environment*, *abstract*, *body*, *thing*, *action*, *emotion*, *spatial*, *plant*, *size*, *qualities*, *cognitive*, *time*, *temperature*, *social*, *animal*, and *question*. Then, the number of concepts C is randomly selected from each of three categories.[4] For each concept, the number of data points P is drawn from the posterior predictive distribution matrix for both vocal and gesture modality.

In $Simulation_4$, the selection of concepts is constrained by three morphological categories – nouns, adjectives, and verbs. The number of concepts C is randomly selected from each category. For each concept, the number of data points P is drawn from the posterior predictive distribution matrix for both vocal and gesture modality.

All simulations were run for each combination of the number of concepts and number of participants, resulting in 36 simulations in total. Simulations were iterated for 100,000 experiments whereby each experiment consisted of a list of concepts and two lists of gesture and vocal expressibility samples for each concept. The simulations were run with a custom-made `Python` script (that is available at the OSF repository).

[4]Note that a category can only be selected for an experiment if there are enough items available.

Statistical analysis Subsequently, for each experiment within a simulation, we performed an independent Student's t-test to determine whether there is a reliable difference between the two modalities. For this purpose, we used the `Pingouin` package (Vallat, 2018) in `Python` which additionally provides the scaled Jeffrey-Zellner-Siow (JZS) Bayes Factor of the alternative hypotheses (i.e., that gesture modality is reliably different from vocal modality). The Cauchy scale factor for computing the Bayes factor was set to 0.5.

For each simulation, we assess the mean and standard deviation of the t-value, p-value, Bayes Factor, and Cohen's d (i.e., the effect size). To assess the robustness of the effect, we additionally calculate the proportions of experiments with inconclusive results, i.e., where p-value > 0.05 and Bayes Factor < 3. Additionally, we use the calculated confidence interval of the difference to assess the mean midpoint of difference for each simulation, together with its standard deviation. Complementary to that, we computed the contrast distribution (i.e., $distribution_{gesture}$ - $distribution_{vocal}$) for one sample experiment in each simulation. This information about confidence interval and distribution contrast will help gauge the actual differences between the distributions of the two modalities.

Results

The extensive descriptive summary of all inferential indicators is displayed in Table 2.

The overall results of the t-tests for all performed simulations show that in the vast number of experiments, we find reliable evidence for accepting the alternative hypothesis, namely that gesture expressibility is different from vocal expressibility. Moreover, since the mean t-value is consistently negative, the concepts have on average lower expressibility in vocalizations than in gestures. The minor proportions of p-values being above the standard threshold of 0.05 (i.e., non-significant) and the minor proportions of the Bayes Factor being below 3 (i.e., anecdotal evidence) further confirm the robustness of the results.

The percentage of inconclusive results differs depending on all varying parameters, as displayed in Figure 2[5]. Despite minor jumps in proportions of anecdotal or no evidence for claiming a difference between gesture and vocal expressibility, we do not find convincing evidence that some types of simulated experiments induce a bias for favoring gesture modality over vocal. We see, for instance, that the proportions of inconclusive results in $Simulations_3$ (random category pick) are double the size of inconclusive results in $Simulation_2$ (categories *action–object–emotion*) – suggesting that for the random category picking there is a higher chance for vocal expressibility being equal (if not better) to gesture expressibility. However, all the proportions of inconclusive evidence are so low that it does not constitute a meaningful difference. Moreover, if the number of concepts is

[5]All figures were made using following `Python`-packages: `matplotlib` (Hunter, 2007), `seaborn` (Waskom, 2021), and `joypy` (see Github for documentation)

parametrized at 24, the differences between simulations are further minimized. Similarly, simulations with 12 concepts only have almost identical proportions of results with insufficient evidence for claiming a difference.

Both contrast distributions (see Figure 2) and mean midpoint of a difference (in Table 2) show that on average, the concepts in vocal modality are approximately 0.166 less expressible than in gesture modality (on a scale from 0 to 1). The effect size remains stable across all simulations, with a mean value of 0.56. The mean standard deviation of the effect size is 0.154, but this value does become higher for fewer concepts and fewer participants (suggesting increased uncertainty about the effect size).

Discussion

Summary The contribution of our work is twofold. First, we provide a set of ratings for 207 concepts, assessed for imagined expressibility by German native speakers in three modalities – multimodal, gesture, and vocal. The results show reliable evidence that ratings are generally highest in the multimodal condition, followed by gesture modality, with vocal modality having the lowest expressibility. This data can be of value to any researcher who wishes to make an informed decision on concept selection guided by people's judgment of expressibility.

In Part II, we focused on whether constraints applied to the selection of concepts, together with a number of concepts and a number of participants, can induce bias in finding one modality being better than the other. Firstly, across all simulations, we find evidence of a difference between gesture and vocal modality, with gestures affording higher expressibility than vocalizations. Given the large average values of the Bayes factor, well above a commonly held threshold of BF>3 for evidence of H1, the likelihood of observing the difference is deemed very robust. The percentage of p-values $> .05$ shows a similar picture (though it is less conservative, showing more 'conclusive evidence' for low sample size experiments as compared to the Bayes Factor thresholded criterion).

It is, however, necessary to take into account the magnitude of the gesture vs. vocal differences observed in our simulations. The mean estimated effect is stable around a medium size of Cohen's d = 0.56. We also observe high variability (SD) of this estimate for different simulations. Across the board, our simulations are more conservative than previously observed 'large' effect sizes (Fay et al., 2014). In this respect, our simulations would thus temper the support for a primacy of one modality over the other, and more critically, the theoretical inferences for language evolution (assuming that such inferences are justified or not).

The simulations also allowed us to test how different concept selections can induce bias in the results. Taking into account the differences in proportions of anecdotal or no evidence for alternative models claiming a difference between two modalities, we do not find convincing evidence of induced bias by picking concepts from particular categories as used in previous research (Fay et al., 2014, 2022).

We acknowledge that there is room for expanding the simulations to further assess potential biases. For instance, we ran a sanity check simulation with only one category that we assumed would have the highest overlap in distribution – *emotion*. We found the proportion of anecdotal or no evidence for a difference rising to 68.761%. This suggests that selecting the 'right' categories may be enough to turn the evidence in the opposite direction. Potentially, one could also tune the simulation in such a way that all semantic categories are used, and from each category, one concept is randomly picked. This would ensure the semantic variance within one stimuli list, and reduce the accumulation of concepts from categories that favor one modality over the other.

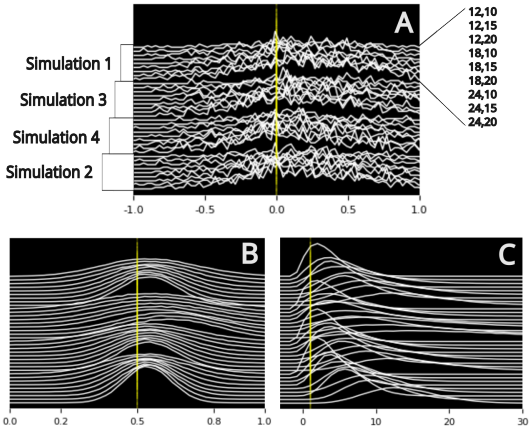

Figure 2: Visualization of output parameters for each simulation type. A: Contrast distribution (yellow line: 0, i.e., no difference). Each distribution represents one sample of a simulation. B: Cohen's d (yellow line = medium-effect size). B: Bayes Factor (yellow line = threshold 3 for anecdotal or no evidence). Note that the values are log-transformed (natural logarithm). For both B and C, each distribution represents one simulation averaged over all 100,000 experiments. All three plots have the same order as displayed in plot A.

Limitations While we believe that the simulations can serve as an insightful (meta)analysis, our approach has methodological and theoretical limitations that should be taken into account.

Methodologically, we address the question of the robustness of expressibility differences in self-reported ratings based on imagined expressibility. We thereby assume that a producer of a concept has a good intuition of not only how well she can express a meaning, but also how likely it is that the guesser perceives the meaning. This does not have to be necessarily the case, as the rater of an online experiment can be oblivious to the feedback from a guesser. Nevertheless, since we introduced the experiment with 'imagine that you have to communicate a meaning to someone else', we be-

Table 2: Summary of the statistical parameters of the Student's t-test.

Simulation (C, P)	t-value		p-value			difference		Cohen's d		Bayes factor			Note
	mean	sd	mean	sd	> 0.05 (in %)	mean	sd	mean	sd	mean (log)	sd (log)	< 3 (in %)	
1 (12, 10)	-4.307	1.473	0.013	0.065	5.213	-0.165	0.053	0.556	0.190	14.853	17.171	10.254	
1 (12, 15)	-5.276	1.694	0.006	0.044	2.247	-0.165	0.050	0.556	0.178	24.986	27.484	4.857	
1 (12, 20)	-6.087	1.883	0.003	0.034	1.266	-0.165	0.048	0.556	0.172	30.352	32.615	2.769	
1 (18, 10)	-5.235	1.448	0.002	0.023	0.978	-0.165	0.043	0.552	0.153	19.878	22.255	2.605	random pick, no constraint
1 (18, 15)	-6.415	1.660	0.001	0.012	0.287	-0.165	0.040	0.552	0.143	30.535	33.030	0.848	
1 (18, 20)	-7.412	1.844	0.000	0.008	0.110	-0.166	0.038	0.552	0.137	41.440	43.919	0.348	
1 (24, 10)	-6.026	1.432	0.000	0.007	0.137	-0.165	0.037	0.550	0.131	23.857	26.323	0.585	
1 (24, 15)	-7.379	1.628	0.000	0.003	0.025	-0.165	0.034	0.550	0.121	35.769	38.269	0.122	
1 (24, 20)	-8.510	1.824	0.000	0.002	0.010	-0.165	0.033	0.549	0.118	49.440	51.930	0.050	
2 (12, 10)	-4.177	1.303	0.010	0.051	4.251	-0.161	0.047	0.539	0.168	12.802	15.161	9.330	
2 (12, 15)	-5.114	1.470	0.004	0.031	1.481	-0.161	0.043	0.539	0.155	17.326	19.412	3.692	
2 (12, 20)	-5.893	1.619	0.002	0.021	0.714	-0.160	0.041	0.538	0.148	23.069	25.490	1.880	
2 (18, 10)	-5.090	1.229	0.001	0.012	0.481	-0.161	0.037	0.536	0.130	18.012	20.512	1.742	random pick from
2 (18, 15)	-6.214	1.368	0.000	0.005	0.089	-0.160	0.033	0.535	0.118	21.925	24.391	0.355	three determined semantic categories
2 (18, 20)	-7.191	1.497	0.000	0.002	0.010	-0.161	0.031	0.536	0.112	29.107	31.549	0.106	*action–object–emotion*
2 (24, 10)	-5.942	1.175	0.000	0.002	0.025	-0.161	0.030	0.542	0.107	17.440	19.511	0.139	
2 (24, 15)	-7.273	1.299	0.000	0.000	0.001	-0.161	0.027	0.542	0.097	24.939	27.190	0.011	
2 (24, 20)	-8.391	1.421	0.000	0.000	0.000	-0.160	0.025	0.542	0.092	36.022	38.374	0.001	
3 (12, 10)	-4.450	1.681	0.017	0.078	6.539	-0.166	0.059	0.575	0.216	18.201	20.669	11.712	
3 (12, 15)	-5.452	1.960	0.009	0.058	3.216	-0.166	0.056	0.575	0.206	28.787	31.063	6.273	
3 (12, 20)	-6.294	2.207	0.005	0.045	1.984	-0.166	0.055	0.575	0.201	42.222	44.688	3.954	
3 (18, 10)	-5.626	1.881	0.005	0.039	1.998	-0.173	0.054	0.593	0.198	25.442	27.928	4.260	random pick from three random categories
3 (18, 15)	-6.903	2.219	0.002	0.026	0.806	-0.174	0.052	0.594	0.191	41.785	44.284	1.944	
3 (18, 20)	-7.964	2.496	0.001	0.019	0.417	-0.173	0.050	0.594	0.186	52.464	54.724	1.103	
3 (24, 10)	-6.875	1.955	0.001	0.013	0.328	-0.183	0.048	0.628	0.179	32.830	35.280	0.971	
3 (24, 15)	-8.407	2.318	0.000	0.007	0.110	-0.183	0.046	0.627	0.173	45.313	47.810	0.347	
3 (24, 20)	-9.736	2.623	0.000	0.006	0.051	-0.183	0.045	0.628	0.169	61.387	63.760	0.148	
4 (12, 10)	-4.279	1.482	0.014	0.065	5.524	-0.165	0.053	0.553	0.191	16.485	18.767	10.774	
4 (12, 15)	-5.243	1.720	0.006	0.047	2.501	-0.165	0.050	0.553	0.181	28.700	31.200	5.258	
4 (12, 20)	-6.057	1.911	0.004	0.036	1.326	-0.165	0.048	0.553	0.174	31.862	34.308	2.984	
4 (18, 10)	-5.196	1.461	0.003	0.026	1.129	-0.165	0.043	0.548	0.154	25.245	27.745	2.971	random pick from
4 (18, 15)	-6.387	1.678	0.001	0.013	0.351	-0.165	0.040	0.550	0.144	32.770	35.251	0.967	three determined morphological categories
4 (18, 20)	-7.363	1.872	0.000	0.009	0.130	-0.165	0.039	0.549	0.140	36.975	39.402	0.448	*noun–adjective–verb*
4 (24, 10)	-5.983	1.441	0.000	0.009	0.172	-0.165	0.037	0.546	0.132	24.942	27.299	0.631	
4 (24, 15)	-7.333	1.650	0.000	0.005	0.035	-0.165	0.035	0.547	0.123	32.783	35.151	0.152	
4 (24, 20)	-8.464	1.836	0.000	0.003	0.017	-0.165	0.033	0.546	0.118	45.639	48.018	0.057	

lieve that there is a good chance that the rater might indeed imagine herself in the guessing role, and rate the expressibility based on: a) how transparent does it feel for her, but also b) how transparent this might look/sound for others. Even if participants might have been overconfident without any external feedback, the data would be inflated uniformly. Moreover, note that we do not operate with raw means but with modeled posterior estimates that can partially cover potential inflation in the data. Nevertheless, we are currently running an experiment to validate these ratings and to address the precise relationship between self-reported expressibility and transparency assessed via production and perception accuracy.

Secondly, there are other ways to statistically analyze the simulations with different assumptions. We draw based on the posterior predictive distribution of the rating data that are not necessarily or optimally Gaussian distributed but perform a test that assumes Gaussian distribution. A case could be made to produce inferential statistics that take into account a possible non-normal distribution (e.g., non-parametric tests). To overcome a possible oversight having to do with non-normal distributions, we also provide the contrastive distributions (together with the mean midpoint of difference) which show that the increase in expressibility from gesture to vocal modality tends to be around 0.166.

Thirdly, the definition of the semantic category is somewhat arbitrary. The top-down approach could be substituted by more data-driven ways of finding clusters of concepts that group together based on some other parameters (e.g., concreteness ratings).

Finally, note that by focusing on some aspects of vocal-gesture differences, such as expressibility, we sidetrack other relevant aspects (e.g., required motor or cognitive energy/effort) that could make the difference between these two modalities more or less important for understanding language evolution.

Theoretical implications At last, we also suggest some theoretical caveats to this type of research that warn against drawing strong inferences from supposed differences without further theoretical specification when and why such a difference would matter for some larger research questions such as how language evolved in humans. For example, we should wonder how much difference in expressibility makes a difference, such that one modality is favored in early communication over another – if both modalities were to some extent useful and allowed for flexible exploitation of affordances, the differences in expressibility are rather marginal. Critically, then, the theoretical import for language evolution of which modality is more expressive is itself a matter of discussion. The question of how different modalities offer different affordances can inform many other overarching research questions (e.g., why and when humans combine modalities, and what concepts solicit expression in one modality rather than another).

Conclusion Our simulations indicate that gestures tend to afford more expressibility (as judged by participants) than vocalizations, regardless of the chosen parameters for a stimuli list. However, vocalizations share a lot of expressibility with gestures, and the difference in expressibility between gesture and vocal modality might be relatively marginal to draw meaningful inferences about early language evolution.

Acknowledgments

This work has been supported by DFG grants FU 791/9-1, CW 10/1-1 and PO 2841/1-1. WP is funded by a VENI grant (VI.Veni 0.201G.047: PI Wim Pouw). We would like to thank the participants of the study.

References

Bohn, M., Kachel, G., & Tomasello, M. (2019). Young children spontaneously recreate core properties of language in a new modality. *Proceedings of the National Academy of Sciences*, *116*(51), 26072–26077.

Bürkner, P.-C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, *80*, 1–28.

Cooperrider, K. (2020). *If language began in the hands, why did it ever leave?* Aeon. Retrieved from https://aeon.co/essays/if-language-began-in-the\-hands-why-did-it-ever-leave

Ćwiek, A., Fuchs, S., Draxler, C., Asu, E. L., Dediu, D., Hiovain, K., … others (2021). Novel vocalizations are understood across cultures. *Scientific Reports*, *11*(1), 10108.

Dezeure, R., Bühlmann, P., Meier, L., & Meinshausen, N. (2015). High-dimensional inference: confidence intervals, p-values and r-software hdi. *Statistical science*, 533–558.

Fay, N., Lister, C. J., Ellison, T. M., & Goldin-Meadow, S. (2014). Creating a communication system from scratch: gesture beats vocalization hands down. *Frontiers in Psychology*, *5*, 354.

Fay, N., Walker, B., Ellison, T. M., Blundell, Z., De Kleine, N., Garde, M., … Goldin-Meadow, S. (2022). Gesture is the primary modality for language creation. *Proceedings of the Royal Society B*, *289*(1970), 20220066.

François, A. (2022). Lexical tectonics: Mapping structural change in patterns of lexification. *Zeitschrift für Sprachwissenschaft*, *41*(1), 89–123.

Fuchs, S., & Ćwiek, A. (2022). Sounds full of meaning and the evolution of language. *Acoustics Today*, 43–51.

Gabry, J. (2021). *cmdstanr: R interface to'cmdstan'*.

Gardner, R. A., & Gardner, B. T. (1969). Teaching sign language to a chimpanzee: A standardized system of gestures provides a means of two-way communication with a chimpanzee. *Science*, *165*(3894), 664–672.

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2024). rstanarm: Bayesian applied regression modeling via stan. *R package version 2.32.1*, *2*(1).

Hewes, G. W. (1976). The current status of the gestural theory of language origin. *Annals of the New York Academy of Sciences*, *280*(1), 482–504.

Hoeschele, M., Wagner, B., & Mann, D. C. (2023). Lessons learned in animal acoustic cognition through comparisons with humans. *Animal Cognition*, *26*(1), 97–116.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. doi: 10.1109/MCSE.2007.55

Kay, M. (2020). *tidybayes: Tidy data and geoms for bayesian models. r package version 2.1. 1.*

Liebal, K., Slocombe, K. E., & Waller, B. M. (2022). The language void 10 years on: multimodal primate communication research is still uncommon. *Ethology Ecology & Evolution*, *34*(3), 274–287.

Lievers, F. S., & Winter, B. (2018). Sensory language across lexical categories. *Lingua*, *204*, 45–61.

Macuch Silva, V., Holler, J., Ozyurek, A., & Roberts, S. G. (2020). Multimodality and the origin of a novel communication system in face-to-face interaction. *Royal Society open science*, *7*(1), 182056.

Matejka, J., Glueck, M., Grossman, T., & Fitzmaurice, G. (2016). The effect of visual appearance on the performance of continuous sliders and visual analogue scales. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5421–5432).

Morey, R. D., Rouder, J. N., Jamil, T., & Morey, M. R. D. (2023). *Package 'bayesfactor'*. Retrieved from https://www.icesi.co/CRAN/web/packages/Bayes\Factor/BayesFactor.pdf

Nölle, J. (2021). *How language adapts to the environment: An evolutionary, experimental approach* (Unpublished doctoral dissertation). The University of Edinburgh.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., … Lindeløv, J. K. (2019). Psychopy2: Experiments in behavior made easy. *Behavior research methods*, *51*, 195–203.

Perlman, M., & Cain, A. A. (2014). Iconicity in vocalization, comparisons with gesture, and implications for theories on the evolution of language. *Gesture*, *14*(3), 320–350.

R Core Team. (2023). *R: A language and environment for statistical computing*. Vienna, Austria.

Robinson, D. (2014). broom: An r package for converting statistical analysis objects into tidy data frames. *arXiv preprint arXiv:1412.3565*.

Slocombe, K. E., Waller, B. M., & Liebal, K. (2011). The language void: the need for multimodality in primate communication research. *Animal Behaviour*, *81*(5), 919–924.

Tadmor, U. (2009). Loanwords in the world's languages: Findings and results. *Loanwords in the world's languages: A Comparative Handbook*, *55*, 75.

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

Vallat, R. (2018). Pingouin: statistics in python. *J. Open Source Softw.*, *3*(31), 1026.

Vehtari, A., Gelman, A., Gabry, J., & Yao, Y. (2021). Package 'loo'. *Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models*.

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. Retrieved from https://doi.org/10.21105/joss.03021 doi: 10.21105/joss.03021

Wegdell, F., Hammerschmidt, K., & Fischer, J. (2019). Conserved alarm calls but rapid auditory learning in monkey

responses to novel flying objects. *Nature ecology & evolution*, *3*(7), 1039–1042.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020.

Wickham, H., & Wickham, M. H. (2017). *Package tidyverse.* World Health Organization Geneva, Switzerland.

Winter, B., Lupyan, G., Perry, L. K., Dingemanse, M., & Perlman, M. (2023). Iconicity ratings for 14,000+ English words. *Behavior research methods*, 1–16.

Zhou, H., van der Ham, S., de Boer, B., Bogaerts, L., & Raviv, L. (2024). Modality and stimulus effects on distributional statistical learning: Sound vs. sight, time vs. space. *PsyArXiv Preprints*.