'Big' sounds bigger in more widely spoken languages

Shiri Lev-Ari*, Ivet Kancheva, Louise Marston, Hannah Morris, Teah Swingler, & Madina Zaynudinova

Dept. of Psychology, Royal Holloway, University of London, Egham, UK

Key words: Sound symbolism; Community size; Language evolution; Communication

*Address correspondence to:

Shiri Lev-Ari

Dept. of Psychology

Royal Holloway, University of London

Egham Hill, Egham TW20 0EX

shiri.lev-ari@rhul.ac.uk

Abstract

Larger communities face more communication barriers. We propose that languages spoken by larger communities adapt and overcome these greater barriers by increasing their reliance on sound symbolism, as sound symbolism can facilitate communication. To test whether widely-spoken languages are more sound symbolic, participants listened to recordings of the words *big* and *small* in widely spoken and less common languages and guessed their meanings. Accuracy was higher for words from widely spoken languages providing evidence that widely spoken languages harbor more sound symbolism. Preliminary results also suggest that widely spoken languages rely on different sound symbolic patterns than less common languages. Community size can thus shape linguistic forms and influence the tools that languages use to facilitate communication.

If asked to guess whether the foreign word *badag* means 'big' or 'small', you might guess that it means 'big', and you would be right. In contrast, if you were asked to guess what the foreign word *ndu* means, you might guess that it means 'small', even though it means 'big' as well. We ask whether you were correct in the former but wrong in the latter because *badag* is in Sundanese, a language estimated to have about 32 million speakers (Eberhard et al., 2020) whereas *ndu* is in Yele, a language estimated to have only 5,000 speakers (Eberhard et al., 2020). It is well known that sounds are sometimes associated with certain meanings, a phenomenon called *sound symbolism*[1]. Sound symbolism has been argued to facilitate language acquisition and language processing (e.g., Imai et al., 2008; Kantartzis et al., 2011; Meteyard et al., 2015; Vinson et al., 2015). This paper tests whether widely spoken languages rely on this tool more than less-common languages because these languages are under greater communicative pressure to be transparent. By focusing on sound symbolism the paper will not only investigate whether and how languages adapt to the communicative needs of their communities but will also examine the role that sound symbolism plays in language.

Languages are spoken in different social environments. These environments impose different communicative challenges. For example, larger communities might have less shared knowledge, greater input variability, and greater difficulty of converging on a shared system. Recent research shows that languages adapt to their social environment (Lupyan & Dale, 2016). In particular, both correlational and experimental studies have found that languages spoken by

[1] In this paper, whenever we refer to sound symbolism, we refer to non-arbitrary associations between sound and meaning that are not language-specific.

larger communities have a simpler and more systematic[2] grammar (Lupyan & Dale, 2010; Raviv et al., 2019). It has been proposed that the reason that larger communities develop languages that are more systematic is because the greater communicative difficulties that they encounter pressure the languages to adapt and become easier for learning and communication. Indeed, Raviv et al. (2019) found that larger groups had greater input variability, a feature that can burden learning and communication. Furthermore, they found that greater input variability at each time point predicted greater increase in systematicity at the next time point. This finding suggests that systematicity rose as a way to overcome the challenge of input variability. Here we test whether sound symbolism is also used by larger communities as a tool to overcome their greater communicative difficulties.

Sound symbolic patterns can facilitate communication because they are based on universal cognitive biases, and they therefore do not rely on shared cultural or linguistic knowledge, prior exposure, or high proficiency in the language. There are several non-mutually-exclusive theories regarding the basis of sound symbolism, including non-linguistic statistical correspondences in the world (e.g., larger objects emitting sounds at lower frequencies as they move or fall), shapes of the articulators during production, and shared properties that might also lead to shared neural correlates in processing (Sidhu & Pexman, 2018). While the basis of sound symbolism is debated, there is evidence that sound symbolism facilitates language learning. For example, three-year-old children learn novel verbs better when the verbs are sound symbolic rather than neutral or sound like their antonyms (Imai et al., 2008; Kantartzis et

[2] Systematicity here refers to consistent mappings between word parts and meanings. It is measured by the correlation between string distances and meaning distances.

al., 2011) and similar results have been obtained with adults (Nielsen & Rendall, 2012). Furthermore, an examination of sound symbolism in English and Spanish shows that words that are acquired at an earlier age are more likely to be sound symbolic (Monaghan et al., 2014; Perry et al., 2015; 2017), and BSL signs acquired earlier are more likely to be iconic (Thompson et al., 2012), supposedly because of iconicity's facilitative role in learning. The facilitative role of sound symbolism and iconicity in general seems to not end in acquisition but also extend to processing, at least in some tasks and under certain circumstances. Thus, people are faster to respond to iconic word sin a lexical decision task (Sidhu, Vigliocco, & Pexman, 2020), individuals with aphasia are faster to read aloud and respond in an auditory lexical decision task to words that are sound symbolic vs neutral (Meteyard et al., 2015), and signers are faster to produce and process iconic signs (Vinson et al., 2015). Thus, while the meaning of sound symbolic words cannot simply be read off the word out of context, their fit with cognitive biases helps guide the listener when interpreting the word, increasing their likelihood of guessing the correct meaning in context, as well as help facilitate learning the word for future use.

Prior research then shows that sound symbolism and iconicity more generally facilitate language acquisition and processing. Prior research also suggests that languages spoken by more people adapt to become easier to use in order to facilitate communication across large communities. The current study therefore tests whether widely spoken languages exploit sound symbolism in order to facilitate communication and overcome their communicative challenges. The hypothesis that widely spoken languages should be more sound symbolic, and therefore more transparent, to facilitate communication is in line with recent research on facial expressions. Such research shows that more heterogeneous communities, which also face

greater communicative challenges, display more exaggerated facial expressions that are more transparent and therefore better understood even by outsiders (Rychlowska et al., 2015; Wood et al., 2016).

As a first test of our hypothesis, that languages that are spoken more widely are more sound symbolic, we presented participants with audio recordings of the words meaning *big* and *small* in languages spoken by particularly large communities and languages spoken by particularly small communities. Participants guessed whether each word meant 'big' or 'small'. We predicted that participants would be better at guessing the meanings of the words that were taken from languages spoken by larger communities. We decided to focus on the words *big* and *small* because there are well-known associations between certain sounds and size. People associate high front vowels with small size and low back vowels with large size (e.g., Newman, 1933; Ohtake & Haryu, 2013; Parise & Spencer, 2012; Peña et al., 2011; Sapir, 1929; Tarte, 1975; Tarte & Barritt, 1971; Thompson & Estes, 2011) and surveys of size words in natural languages also uncovered similar patterns, although sometimes only for one of the two meanings or only in certain word positions (Blasi et al., 2016; Haynie, Bowern & LaPalobara, 2014; Winter & Perlamn, 2021). Therefore, in addition to testing whether people are better at guessing the meaning of words in more widely spoken languages, we tested whether individuals are more likely to guess 'small' when a word has high front vowels, and more likely to guess 'big' when it has low back vowels, and whether the words for *big* and *small* in the more widely spoken languages are more likely to exhibit the correspondence between vowels and size.

Study

Method

Participants. One-hundred-twenty-eight individuals (F=74) with normal hearing participated in an online experiment. Ninety-five of them were native speakers of English. Others reported their native languages as Russian (N=8), Bulgarian (N=4), Urdu (N=4), German (N=3), Polish (N=3), Spanish (N=2), French (N=2), Farsi, Hungarian, Luxembourgish, Portuguese, Romanian, Somali, and Arabic-Polish (1 each). We aimed for a sample size of 120 based on Bankieris and Simner (2015) while adjusting their sample size to the fact that our design was within-participants and our preference to recruit a larger sample size than theirs as not all of their comparisons reached significance.

Stimuli. We selected 20 languages spoken by millions of people, avoiding familiar European languages (Median=79.2m; range: 24.7m-1.1billion) and 20 languages spoken by only hundreds or thousands of people (Median=2,870; range: 200-328,080; See Table 1). All translations were gathered from sources providing Swadesh lists for those languages and are provided in Appendix A. We generated audio files for the words *big* and *small* for each of those languages using text-to-speech synthesizers. Because there are no text-to-speech synthesizers for the less common languages, we generated words for all languages using an Esperanto speech synthesizer (https://parol.martinrue.com/), as it is relatively neutral. When needed, we collapsed over similar phonemes to fit the Esperanto phoneme inventory. For example, /ɛ/ and /e/ were both produced as /e/, /i/ and /ɪ/ were both produced as /i/ etc. Seven of the words (four from widely spoken languages and three from less common languages) contained central

vowels, which Esperanto does not have. These were therefore generated with a Romanian synthesizer (https://texttospeech.io/text-to-mp3-online), because Romanian includes these sounds and is not among the tested languages.

Table 1. List of languages used in study

Language	**Language family**	**Number of Speakers**	**Population classification**
Mandarin Chinese	Sino-Tibetan	1,119,961,120	Large
Hindi	Indo-European>Indo-Aryan	600,485,970	Large
Standard Arabic	Afro-Asiatic>Semitic	346,922,980	Large
Russian	Indo-European>Balto-Slavic	258,034,160	Large
Indonesian	Austronesian	198,990,530	Large
Japanese	Japonic	126,379,110	Large
Telugu	Dravidian	95,581,000	Large
Turkish	Turkic	88,101,920	Large
Tamil	Dravidian	85,456,100	Large
Korean	Koreanic	81,520,400	Large
Vietnamese	Austroasiatic>Vietic	76,843,160	Large
Hausa	Afro-Asiatic>Chadic	74,930,300	Large
Swahili	Niger-Congo>Bantu	69,195,410	Large
Kannada	Dravidian	58,644,310	Large
Amharic	Afro-Asiatic>Semitic	57,445,260	Large
Burmese	Sino-Tibetian> Lolo-Burmese	42,954,860	Large
Polish	Indo-European>Balto-Slavic	40,646,160	Large
Sundanese	Austronesian>Malayo-Polynesian	32,400,000	Large
Zulu	Niger-Congo>Bantu	27,770,100	Large
Nepali	Indo-European>Indo-Aryan	24,720,300	Large
Icelandic	Indo-European> Germanic	328,080	Small

Papel	Niger-Congo> Atlantic Congo	173,500	Small
Daasanach	Afro-Asiatic>Cushitic	66,630	Small
Lepcha	Sino-Tibetan	57,930	Small
Korwa	Austroasiatic>Munda	28,500	Small
Badyara	Niger-Congo>Atlantic Congo	20,510	Small
Nduga	Trans New-Guinea> West	10,000	Small
Orokolo	Trans New-Guinea> Eleman	7,500	Small
Yele	unclassified	5,000	Small
Affetti / Afitti	Nilo-Saharan> Eastern Sudanic	4,000	Small
Walman/Valman	Torricelli	1,740	Small
Manx	Indo-European>Celtic	1,660	Small
Talodi	Niger-Congo> Kordofanian	1,500	Small
Hunzib	Northeast Caucasian> Tsezic	1,420	Small
Alamblak	Sepik	1,000	Small
Chambri	Ramu-Lower Sepik	800	Small
Juwal	Torricelli	700	Small
Eritai	Lakes Plain	530	Small
Zenaga	Berber	200	Small
Fongoro	Nilo-Saharan>Central Sudanic	50[3]	Small

Procedure. Participants heard the words in random order. The task was self-paced and participants could replay an audio file up to three times if they wished. For each word, they guessed whether the word meant 'big' or 'small'. Participants were instructed that if they knew

[3] The number of speakers of Fonogro is lower than 50. We adjusted it to 50 in line with Lupyan & Dale's (2010) approach.

the word or the word sounded similar to a word they knew, they should select the "I recognize this word" option.

Results

All stimuli, data, and analysis scripts are available at: https://osf.io/7p32b/. Before analyzing the results we excluded all responses that were shorter than the duration of the audio file. This led to the exclusion of 135 responses (1.3%). The majority of these too fast responses (108/135=80%) belonged to three participants. We excluded the remaining responses of these participants as well. Additionally, as we examined the results, we realized that the recordings of two of the words were very similar to English words /lɪtɪtl/ which sounded quite similar to *little*, and /nini/ whose recording was ambiguous and sounded closer to *mini*. We therefore excluded these words from analysis.

To test whether participants can better guess the meaning of words from widely spoken languages, we conducted a logistic mixed effects regression using the lme4 package 1.1.27.1 (Bates et al., 2016) in R 4.0.2 (R core team, 2020). We wrangled the data with the tidyverse package 1.3.0 (Wickham et al., 2019) and plotted our results with the effects package 4.1.4 (Fox & Weisberg, 2019). The model included Community Size (large, small) as a fixed effect, Participants and Items as random effects, and accuracy as the dependent measure. Results indicated that, as predicted, participants were better at guessing word meanings in languages spoken by many vs few people (59% vs 52%; β=-0.3, SE=0.15, z=-2, $p<0.05$; See Fig 1a and Appendix B for the full table of results). We also conducted an exploratory analysis using the (log) number of speakers from Ethnologue (Eberhard et al., 2020) as a predictor rather than the

categorical variable of Community Size. Results showed an effect of (log) number of speakers (β=0.03, SE=0.01, z=2.1, p<0.04; Fig 1b). The logarithmic fit suggests that, as is common with community size effects, the effect of number of speakers is larger for languages with smaller community size, and that after a language reaches a certain size, a further increase in community size does not increase its degree of sound symbolism as much. That said, future studies should sample more uniformly across the entire range of population size to better assess the logarithmic vs linear nature of the effect.

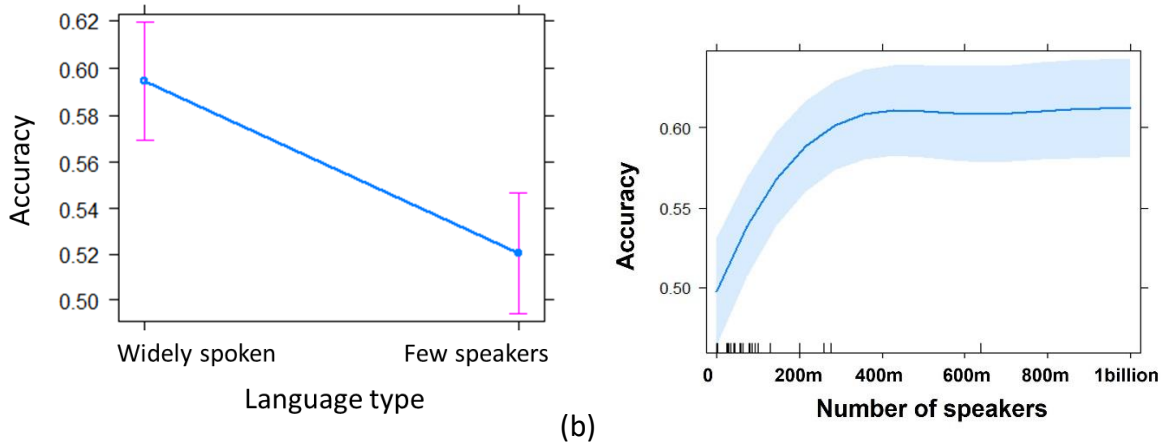

Figure 1. Accuracy of guessing word meanings in a language as dependent on the size of the community speaking the language. Figure 1a shows the analysis using the categorical variable Community Size, Figure 1b shows the analysis using the log number of speakers.

Next, we examined whether participants relied on the established association between front/back vowels and small/large size to make their judgments (e.g., Newman, 1933; Ohtake & Haryu, 2013; Parise & Spencer, 2012; Peña et al., 2011; Sapir, 1929; Tarte, 1975; Tarte & Barritt, 1971; Thompson & Estes, 2011; Winter & Perlman, 2021). We coded the vowel frontness of each word by granting a point for every front vowel (/i/ and /e/), and subtracting a point for

every back vowel (/a/, /o/, and /u/). Central vowels were granted 0 points. Vowel Frontness scores ranged from -3 to 4 (M=-0.34). We conducted a logistic mixed effects model with Vowel Frontness as a fixed effect, and response (small=1, big=0) as the dependent variable. The random structure included intercepts for Participants and Items. Results revealed that participants were more likely to guess that a word means "small" the more front vs back vowels it had (β=0.16, SE=0.05 z=3.09, p<0.01[4]; Fig 2). A further analysis revealed that the higher a word's Vowel Frontness score, the more likely the word is to mean 'small' (β=1.15, SE=0.46, z=2.5, p<0.02), but, in contrast to our prediction, the more widely spoken languages were not more likely than the less common languages to exhibit this vowel-size association (p>0.2). This indicates that widely spoken languages rely on different sound symbolic cues than the vowel-size association to increase their transparency vs the less-common languages. As an anecdote, the words *big* and *small* in English show the opposite pattern than would be expected by this association, though an analysis of the entire size vocabulary in English (e.g., tiny, huge) shows that the association between front vowels (in particular /i/and /ɪ/) and small size, and back vowels, (in particular /ɑ/) and large size is evident in English as well (Winter & Perlman, 2021).

[4] We also ran an analysis that examined the role of front and back vowels separately. The analysis revealed that the more back vowels there were, the less likely were participants to select 'small' (β=-0.23, SE=0.10, z=-2.16, p=0.0306). The effect of number of front vowels was not significant on its own although, numerically, it went in the predicted opposite direction, namely, more front vowels numerically increased the likelihood of selecting "small".

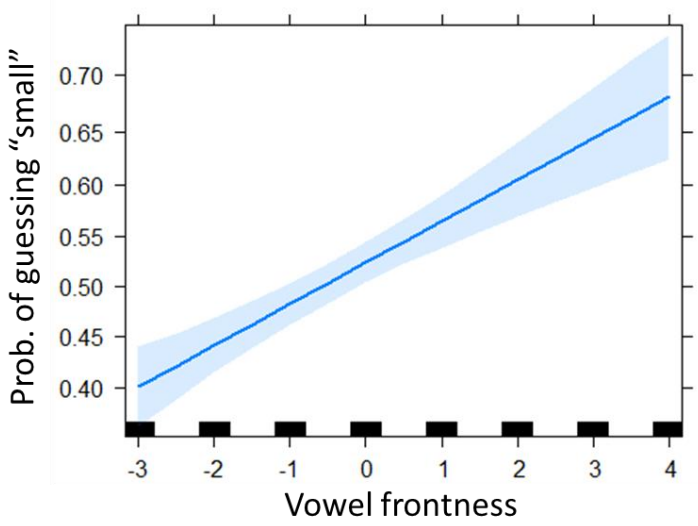

Figure 2. The effect of Vowel Frontness (number of front vowels – number of back vowels) on participants' likelihood of responding 'small' (vs 'big').

We carried out exploratory analyses to reveal which other cues participants relied on. Correlations between the appearance of certain consonants in the words and participants' guesses suggested the occurrence of /b/ (r=-0.61) and /g/ (r=-0.42) in the word increased "big" responses and the occurrence of /n/ (r=0.31) increased "small" responses. Re-running the analysis of Vowel Frontness while adding predictors for occurrence of these three phonemes showed the previously found effect of Vowel Frontness (β=0.09, SE=0.04, z=2.48, p<0.02) as well as negative effects of the occurrence of /b/ and /g/ on responding "small" (/b/: β=-0.95, SE=0.14, z=-6.76, p<0.001; /g/: β=-1.04, SE=0.18, z=-5.61, p<0.001). The effect of /n/ did not reach significance (p=0.06).

This raises the question of whether words from widely spoken languages were better guessed because they were more likely to rely on these consonant-size associations. We do not have enough data to investigate this hypothesis statistically, but the numerical patterns are in line

with this hypothesis: As illustrated in Figure 3, in widely spoken language, /b/ occurred exclusively in words meaning *big* whereas /n/ occurred almost exclusively in words meaning *small*. In contrast, in less common languages, /b/ was equally likely to occur in words meaning *big* and *small*, and the distribution of /n/ was also less skewed. The phoneme /g/ was quite rare across all languages. Together, these preliminary data suggest that whereas less common languages rely mostly on vowel-size associations (since they did exhibit the established vowel-size correspondence but they do not seem to exhibit a consonant-size associations), widely spoken languages also rely on consonant-size associations.

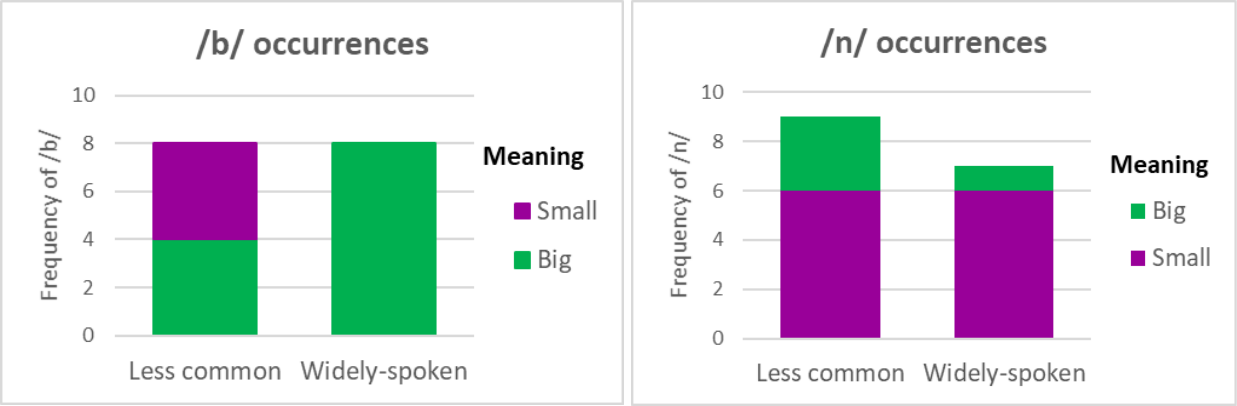

Figure 3. Frequency of /b/ and /n/ in words meaning *big* and *small* in widely spoken and less common languages

Despite these suggestive results, since all our participants spoke English, we wondered whether participants' knowledge of the word "big" led to the associations of /b/ and /g/ rather than a true sound-symbolic pattern. As a preliminary exploration, we collected the words *big* and *small* from all languages featured in the 150 most common languages according to Ethnologue (Eberhard et al., 2020). We filtered the words to avoid cognates: When words from related

languages had >50% phonemic overlap, we only kept the one from the language with the higher number of speakers. This led to a list of 59 languages (the 20 original widely spoken languages and 39 additional languages) that each contributed two words: *big* and *small*. We then conducted Wilcoxon signed-rank tests with the Pratt method for dealing with zero values comparing the occurrence of each of the phonemes /b/, /g/, and /n/ in the words meaning "small" vs "big". Since the comparison was within languages, it controls for the frequency of the phonemes in the language, since phoneme frequency should equally influence the phoneme's likelihood of appearing in the word meaning "big" and the word meaning "small". All three Wilcoxon signed-rank tests showed effects in the predicted direction (/b/: 18 occurrences in words meaning "big" vs 1 in words meaning "small" ,$W=50.5$, $p<0.001$; /g/: 12 vs 3, $W=156$, $p<0.03$; /n/: 11 vs 26, $W=240$, $p<0.02$). Together, there is some preliminary evidence to suggest that /b/, /g/, and /n/ are associated with large and small size, respectively and that widely spoken languages might capitalize on this association more. It should be noted that several earlier studies have also found an association between voiced consonants and large size (Klinck, 2000; Monaghan & Fletcher, 2019) including, in particular, /b/ and /g/ (Westbury et al., 2018).

Discussion

This study shows that widely spoken languages use more sound symbolic size vocabulary than languages spoken by few people. We propose that this is driven by the need to overcome the greater communicative challenges in larger communities. Interestingly, the study also suggests that community size might not only influence the degree to which languages rely on sound symbolism but also which sound symbolic patterns they exploit.

The hypothesis that this study set out to test is that widely spoken languages are more sound symbolic than languages spoken by fewer people, but the task itself focused solely on the words *big* and *small*. One may therefore wonder how generalizable the finding is to other words and dimensions. While a definite answer can only be provided with further investigation, our choice of words was rather conservative with the potential of under-estimating the effect. The words *big* and *small* are considered basic terms and were taken from the Swadesh list. Basic terms such as these are often assumed to be more stable across time, and this is one of the features that renders them useful for typological research. In other words, these words are assumed to be less vulnerable to changes in social structure. Despite this fact, these words showed sensitivity to community size. One would expect that the effect of community size would be even larger in words that are more malleable. Similarly, if one of the main goals of sound symbolism in large communities is to facilitate communication, one might expect that the effect of community size on degree of sound symbolism might be even stronger in words related to inter-group contact, or to survival. On the other hand, size is a dimension that easily affords sound symbolism. Therefore, it is also possible that meanings that do not afford sound symbolism as easily would exhibit a smaller effect of community structure on sound symbolism. It is also possible that the effect of community size might have been different if instead of examining conventional words we looked at onomatopoeia, as it has been suggested that the latter is more common among less-industrialized societies (e.g., Berlin & O'Neill, 1981). Future studies should further investigate the generalizability of the effect of community size and whether it is more prevalent in some areas of the lexicon vs others.

We propose that more widely spoken language are more sound symbolic because larger communities encounter more communicative obstacles. That said, we did not test the mechanism directly and community size correlates with other variables, such as the number of non-native speakers. Indeed, Lupyan and Dale (2010) interpreted their finding that more widely spoken languages have simpler morphology as stemming from the correlation between community size and proportion of non-native speakers. While a follow up experimental study showed that community size, on its own, can lead to more systematic grammar (Raviv et al., 2019), other studies also found effects of the proportion of non-native speakers on grammar and vocabulary (Bentz et al., 2015; Bentz & Winter, 2014). Larger communities and heterogeneous communities with many non-native speakers encounter many similar communicative obstacles. Therefore, it is quite likely that the proportion of non-native speakers would also exert an influence on the level of sound symbolism. It might therefore be the case that the effect of community size found here encompasses effects of several community properties that all correlate with each other and are all associated with encountering greater communicative challenges. Thus, while this study demonstrates that widely spoken languages are more sound symbolic, at least when it comes to size words, future experimental work should test whether it is indeed the communicative challenges of larger communities that lead to this effect.

Widely spoken language are also more likely to have had contact with English. This could have led to mutual influence which would have made it easier for our participants to guess the meanings. While we avoided familiar European languages and filtered out any words that seemed similar to English words that refer to size, we cannot fully rule out the possibility that

any contact between the languages have led the languages to adapt and become more comprehensible to English speakers.

The findings about a shift in the sound-symbolic patterns that languages utilize as their community of speakers grows are still preliminary but merit further investigation. One of the properties in which languages of larger and smaller communities differ is the greater input variability in larger communities. Phonetically, individuals vary more in their production of vowels than of consonants (Kleinschmidt, 2016). Therefore, vowels would be a less reliable cue in larger communities than smaller communities. This could lead languages spoken by larger communities to rely more on consonants-based patterns than vowel-based patterns. The greater variability in vowel production might even push languages to adapt their phonological inventory in general to increase reliance on consonants vs vowels to increase communicative success. Prior research has provided controversial findings that community size can influence a language's phonological inventory (Hay & Bauer, 2007). Further research should examine whether community size influences the internal structure of the inventory.

If the sound symbolic patterns in widely spoken languages differ from those in less common languages, one may wonder whether the fact that participants were native speakers of widely spoken languages provided an advantage to the widely spoken languages, as participants might have been particularly sensitive to sound symbolic patterns that are similar to those in their own languages. While we cannot rule out this possibility, it requires community structure to have an effect on sound symbolic patterns, yet that these sound symbolic patterns would not be rooted in universal cognitive associations and biases. That is, we propose that more widely

spoken languages increase their reliance on sound symbolism because it facilitates communication as it relies on shared biases. This alternative proposal suggests that community structure influences sound symbolic patterns but that even though the process would occur across disparate widely spoken languages, it would not rely on universal biases, and therefore would be better understood by speakers of languages that exhibit the same patterns. It is unclear what would motivate such a process, but we cannot rule out the possibility that even patterns that are based on cognitive associations and biases might be better understood if you encounter them more oftne in your own language.

The finding that widely spoken languages are more sound symbolic also addresses the question regarding the utility and extent of sound symbolism. Past research has often focused on the utility of sound symbolism for language acquisition by children (e.g., Imai et al., 2008; Monaghan et al., 2014). The greater reliance on sound symbolism in widely spoken languages suggests that it might fulfill other functions as well such as facilitating communication between individuals with limited shared knowledge. In that case, as mentioned earlier, we might expect that, especially in languages spoken by large and heterogeneous communities, sound symbolism would be more common in words relating to survival or intergroup relations. Similarly, in communities with a high proportion of second language learners, sound symbolism might be particularly common not only in words acquired at an early age but in words acquired early by adult learners. Future research should investigate the occurrence of sound symbolism in such semantic domains, and examine whether the semantic domains in which sound symbolism is more prominent vary with properties of the community.

Considering the benefits of sound symbolism, one may wonder why sound symbolism is not more frequent, and why less common languages do not also rely on it as much as well. First, while sound symbolism confers benefits, so does arbitrariness. For example, arbitrariness has been argued to allow generalization (Lupyan & Winter, 2018). Therefore, languages need to weigh the benefits of each type of word. When communicative challenges are greater, the relative importance of facilitating understanding might increase whereas other pressures remain the same, leading to preference for higher level of sound symbolism. Furthermore, even when a feature is unambiguously useful, it might not develop when there is no strong pressure for it. For example, in Raviv et al. (2019) smaller groups developed less systematic languages than larger groups even though systematicity is beneficial, because the smaller groups managed to reach high level of performance even without it whereas the larger ones needed to employ that tool to reach the same level of performance. Thus, smaller groups might not resort as much to sound symbolism because communication can be highly successful even without it.

To conclude the study shows that languages adapt to their communicative needs by modulating their reliance on sound symbolism. It also suggests that community size might influence the types of sound symbolism they employ. It thus shows how languages evolve to facilitate communication in a manner appropriate to their specific social needs.

Refences

Bankieris, K., & Simner, J. (2015). What is the link between synaesthesia and sound symbolism?. *Cognition, 136*, 186-195.

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2016). *lme4: Mixed-effects modeling with R*; 2010. Available at https://cran.r-project.org/web/packages/lme4

Bentz, C., Verkerk, A., Kiela, D., Hill, F., & Buttery, P. (2015). Adaptive communication: Languages with more non-native speakers tend to have fewer word forms. *PLoS One*, *10*(6), e0128254.

Bentz, C., & Winter, B. (2014). Languages with more second language learners tend to lose nominal case. In *Quantifying language dynamics* (pp. 96-124). Brill.

Berlin, B., & O'Neill, J. P. (1981). The pervasiveness of onomatopoeia in Aguaruna and Huambisa bird names. *Journal of Ethnobiology, 1, 2*, 238–261.

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences, 113, 39*, 10818–10823.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). (2020). Ethnologue: Languages of the World. Twenty-third edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Fox, J. & Weisberg, S. (2019). An R Companion to Applied Regression, 3rd Edition. Thousand Oaks, CA

Hay, J. & Bauer, L. (2007) Phoneme inventory size and population size. *Language 83*, 388–400.

Haynie, H., Bowern, C., & LaPalombara, H. (2014). Sound symbolism in the languages of Australia. *PLoS ONE, 9*, 4, e92852.

Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition, 109*, 54-65.

Kantartzis, K., Imai, M., & Kita, S. (2011). Japanese sound-symbolism facilitates word learning in English-speaking children. *Cognitive Science*, *35*(3), 575-586.

Kleinschmidt, D. F. (2016). *Perception in a variable but structured world: The case of speech perception* (Unpublished doctoral dissertation). University of Rochester, NY

Klink, R. R. (2000). Creating brand names with meaning: The use of sound symbolism. *Marketing Letters, 11*, 1, 5-20.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PloS One, 5, 1*, e8559.

Lupyan, G., & Dale, R. (2016). Why are there different languages? The role of adaptation in linguistic diversity. *Trends in cognitive sciences, 20, 9*, 649-660.

Lupyan, G., & Winter, B. (2018). Language is more abstract than you think, or, why aren't languages more iconic?. *Philosophical Transactions of the Royal Society B: Biological Sciences, 373, 1752*, 20170137.

Meteyard, L., Stoppard, E., Snudden, D., Cappa, S. F., & Vigliocco, G. (2015). When semantics aids phonology: A processing advantage for iconic word forms in aphasia. *Neuropsychologia, 76*, 264-275.

Monaghan, P., & Fletcher, M. (2019). Do sound symbolism effects for written words relate to individual phonemes or to phoneme features?. *Language and Cognition, 11, 2*, 235-255.

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language?. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130299.

Newman, S. S. (1933). Further experiments in phonetic symbolism. *The American Journal of Psychology, 45, 1*, 53–75.

Nielsen, A., & Rendall, D. (2012). The source and magnitude of sound-symbolic biases in processing artificial word material and their implications for language learning and transmission. *Language and cognition*, *4*(2), 115-125.

Ohtake, Y., and Haryu, E. (2013). Investigation of the process underpinning vowel size correspondence. *Japanese Psychological Research, 55*, 390–399.

Parise, C. V., & Spence, C. (2012). Audiovisual crossmodal correspondences and sound symbolism: A study using the implicit association test. *Experimental Brain Research, 220*, 319–333

Peña, M., Mehler, J., & Nespor, M. (2011). The role of audiovisual processing in early conceptual development. *Psychological Science, 22*, 1419–1421.

Perry, L. K., Perlman, M., & Lupyan, G. (2015). Iconicity in English and Spanish and its relation to lexical category and age of acquisition. *PloS One, 10, 9*, e0137147.

Perry, L. K., Perlman, M., Winter, B., Massaro, D. W., & Lupyan, G. (2017). Iconicity in the speech of children and adults. *Developmental Science*, e12572

R Core Team. (2020). *A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: http://www.R-project.org

Raviv, L., Meyer, A., & Lev-Ari, S. (2019). Larger communities create more systematic languages. *Proceedings of the Royal Society B, 286*, 1907, 20191262.

Rychlowska, M., Miyamoto, Y., Matsumoto, D., Hess, U., Gilboa-Schechtman, E., Kamble, S., ... & Niedenthal, P. M. (2015). Heterogeneity of long-history migration explains cultural differences in reports of emotional expressivity and the functions of smiles. *Proceedings of the National Academy of Sciences, 112, 19*, E2429-E2436.

Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology, 12*, 225–239

Sidhu, D. M., & Pexman, P. M. (2018). Five mechanisms of sound symbolic association. *Psychonomic bulletin & review, 25, 5*, 1619-1643.

Sidhu, D. M., Vigliocco, G., & Pexman, P. M. (2020). Effects of iconicity in lexical decision. *Language and Cognition, 12, 1*, 164-181.

Tarte, R. D. (1974). Phonetic symbolism in adult native speakers of Czech. *Language and Speech, 17, 1*, 87–94.

Tarte, R. D., & Barritt, L. S. (1971). Phonetic symbolism in adult native speakers of English: Three studies. *Language and Speech, 14, 2*, 158–168.

Thompson, P., and Estes, Z. (2011). Sound symbolic naming of novel objects is a graded function. *Quarterly Journal of Experimental Psychology, 64*, 2392–2404.

Thompson, R. L., Vinson, D. P., Woll, B., & Vigliocco, G. (2012). The road to language learning is iconic: Evidence from British Sign Language. *Psychological science, 23, 12*, 1443-1448.

Vinson, D., Thompson, R. L., Skinner, R., & Vigliocco, G. (2015). A faster path between meaning and form? Iconicity facilitates sign recognition and production in British Sign Language. *Journal of Memory and Language, 82*, 56-85.

Westbury, C., Hollis, G., Sidhu, D. M., & Pexman, P. M. (2018). Weighing up the evidence for sound symbolism: Distributional properties predict cue strength. *Journal of Memory and Language, 99*, 122-150.

Wickham et al., (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4, 43*, 1686,

Winter B. & Perlman M., (2021) Size sound symbolism in the English lexicon", *Glossa: a journal of general linguistics 6, 1*, p.79.

Wood, A., Rychlowska, M., & Niedenthal, P. M. (2016). Heterogeneity of long-history migration predicts emotion recognition accuracy. *Emotion, 16, 4*, 413.

Appendix A – list of words (rendered in English)

Language	**Small**	**Big**
Affetti	kacha	duba
Alamblak	habrienir	bodokam
Amharic	tinis	tilik
Badyara	nentiti	manende
Burmese	se	kshi
Chambri	bapoko	ubo
Daasanach		gudu
Eritai	korakikore	sokukwei
Fongoro	katri	kubo
Hausa	karami	baba
Hindi	chota	bara
Hunzib	ieru	idu
Icelandic		stor
Indonesian	kechil	besar
Japanese	chisai	oki
Juwal	tuwemoremp	peyet
Kannada	chika	doda
Korean	chakta	kuda
Korwa	sani	kad

Lepcha	chubu	atin
Mandarin Chinese	shiau	ta
Manx	beg	mur
Nduga	matiyo	gwok
Nepali	sano	tulo
Orokolo	hekai	eapapo
Papel	ontes	omagi
Polish	mawi	velki
Russian	malinkiy	bolshoy
Standard Arabic	sarir	kabir
Sundanese	litik	badag
Swahili	dogo	kubwa
Talodi	isare	utik
Tamil	chiriya	periya
Telugu	chinna	peda
Turkish	kichik	biyik
Vietnamese	no	lon
Walman	volu	lapo
Yele	tire	ndu
Zenaga	imezigen	ioxshi
Zulu	ntsane	kulu

All the words from the languages with small communities (except for Icelandic) were taken from the Rosetta project, obtained from:

https://archive.org/search.php?query=swadesh+collection%3Arosettaproject&sort=titleSorter&and[]=subject%3A%22Swadesh+List%22

The words for Amharic, Indonesian, Kannada, Nepali, Polish, Russian, Sundanese, Tamil, Telugu, Turkish, Vietnamese were taken from asjp.com:

https://asjp.clld.org/

The words for Standard Arabic, Burmese, Hausa, Hindi, Icelandic, Japanese, Korean, Mandarin, Swahili and Zulu were taken from Wiktionary:

https://en.wiktionary.org/wiki/Appendix:Swadesh_lists

Appendix B – table of results

Results of the confirmatory analysis of the effect of community size on accuracy with Language Community Size is examiend categorically (Widely spoken, Less common)

Random effects				
Variable	Variance	SD		
Participant	0.01	0.12		
Item	0.41	0.64		
Fixed effects				
	β	SE	z	p-value
(intercept)	0.39	0.11	3.68	0.000234
Community size (widely spoken	-0.30	0.15	-2.0	0.0455

Results of an exploratory analysis of the effect of community size on accuracy where Language Community Size as a continous variable (log-transformed)

Random effects				
Variable	Variance	SD		
Participant	0.01	0.12		
Item	0.41	0.64		
Fixed effects				

	β	SE	z	p-value
(intercept)	-0.15	0.20	-0.75	0.4515
Community size (widely spoken	0.03	0.01	2.10	0.0355

Results of the effect of vowel frontness of selection

Random effects				
Variable	Variance	SD	Correlation	
Participant (intercept)	0.03	0.17		
Items (intercept)	0.42	0.65		
Fixed effects				
	β	SE	z	p-value
(intercept)	0.09	0.08	1.17	0.243
Vowel frontness	0.17	0.05	3.09	0.002

A test whether the words themselves reflect the vowel-size association

	β	SE	t	p-value
(intercept)	-0.85	0.33	-2.61	0.0108
Meaning (small)	1.15	0.46	2.50	0.0146

Community size (widely-spoken)	0.25	0.46	0.54	0.5883
Meaning x Community size	-0.75	0.65	-1.15	0.2525