Sound-meaning association biases evidenced across thousands of languages

Damián E. Blasi [*] [†], Soeren Wichmann [‡] [§], Harald Hammarström [†], Peter F. Stadler [¶] [‖] [**], and Morten H. Christiansen [††] [‡‡]

[*]University of Zürich,[†]Max Planck Institute for the Science of Human History,[‡]University of Leiden,[§]Kazan Federal University,[¶]University of Leipzig,[‖]Max Planck Institute for Mathematics in the Sciences,[**]Santa Fe Institute,[††]Cornell University, and [‡‡]Aarhus University

Submitted to Proceedings of the National Academy of Sciences of the United States of America

It is widely assumed that one of the fundamental properties of spoken language is the arbitrary relation between sound and meaning. Some exceptions in the form of non-arbitrary associations have been documented in linguistics, cognitive science, and anthropology but these studies only involved small subsets of the 6000+ languages spoken in the world today. By analyzing word lists covering nearly two-thirds of the world's languages, we demonstrate that a considerable proportion of 100 basic vocabulary items carry strong associations with specific kinds of human speech sounds, occurring persistently across continents and linguistic lineages (linguistic families or isolates). Prominently among these relations we find property words ('small' and *i*, 'full' and *p* or *b*) and body part terms ('tongue' and *l*, 'nose' and *n*). The areal and historical distribution of these associations suggests that they often emerge independently rather than being inherited or borrowed. Our results therefore have important implications for the language sciences, given that non-arbitrary associations have been proposed to play a critical role in the emergence of cross-modal mappings, the acquisition of language, and the evolution of our species' unique communication system.

linguistics | cognitive sciences | sound symbolism | language evolution | iconicity

Although there is substantial debate in the language sciences over how to best characterize the features of spoken language, there is nonetheless a general consensus that the relationship between sound and meaning is largely arbitrary [1, 2, 3]. Plenty of exceptions exist, however, within individual languages. For instance, ideophones—a class of words found in many languages—convey a communicative function (or meaning) through the depiction of sensory imagery [4]. In the Mel language Kisi Kisi (spoken in Sierra Leone) *hábá* means "(human) wobbly, clumsy movement", and *hábá-hábá-hábá* "(human) prolonged, extreme wobbling"; here repetition serves as a way to convey the meaning of intensity. More generally, the resemblance between certain aspects of the acoustic basis of speech and their referents, *iconicity*, is the most researched and well-known case of non-arbitrary associations between sound and meaning [5]. *Systemacity*, in contrast, refers to (statistical) regularities that are common to particular set of words, created by historical contingencies and analogical processes [5]. For example, word-initial *gl-* in English evokes the idea of a visual phenomenon (as in *glare*, *glance*, *glimmer*) [6]. At a larger scale, there is evidence that the phonological properties of whole morphosyntactic classes of words (like verbs and nouns) are distinct in several languages [7].

The evidence of recurring regularities in sound-meaning mappings across multiple languages is considerably more modest, despite its potential importance for fundamental questions about language evolution and the role of basic perceptual biases in cognition. For example, certain shape-sound associations—known as the *bouba-kiki* effect [8, 9, 10]—are believed to rely on the ability that humans (and perhaps also other primate species [11]) have for associating stimuli across different modalities [12]. Other plausible sources of cross-linguistic associations include, for instance, the relationship across many animal species between vocalization frequency and animal size [13], the mimicry of referents via unconscious mouth gesturing [14], and the persistence of vestiges of a conjectured early human language [15].

Experimental studies support the hypothesis that humans are indeed sensitive to such associations. It has been demonstrated several times that paticipants perform above-chance when asked to pair up words with opposite meanings (antonyms) in languages unknown to them [16], and that English speakers might even be able to decide on the concreteness

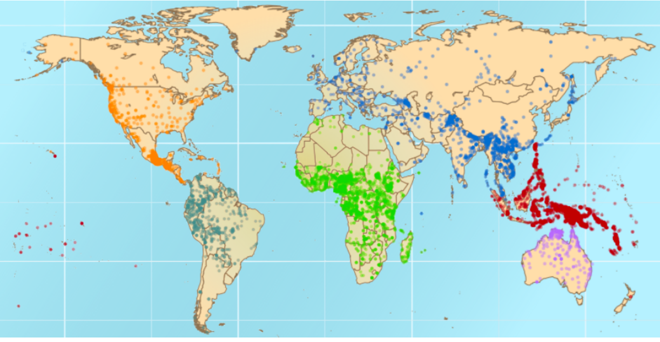


Fig. 1. Geographic distribution of the 6452 word lists from the ASJP database [24]. Colors distinguish different linguistic macro-areas, regions with relatively little or no contact between them (but with much internal contact between their populations). These are North America (orange), South America (dark green), Eurasia (blue), Africa (green), Papua New Guinea and the Pacific Islands (red) and Australia (fuchsia).

Significance

The independence between sound and meaning is believed to be a crucial property of language: across languages, sequences of different sounds are used to express similar concepts (e.g., Russian 'ptitsa', Swahili 'ndege', Japanese 'tori' all mean 'bird'). However, a careful statistical examination of words from nearly two-thirds of the world's languages reveals that unrelated languages very often use (or avoid) the same sounds for specific referents. For instance, words for tongue tend to have l or u, 'round' often appears with r, and 'small' with i. These striking similarities call for a re-examination of the fundamental assumption of the arbitrariness of the sign.

Reserved for Publication Footnotes

of words from languages to which they have not been exposed [17]. However, this evidence for non-arbitrary sound-meaning associations pertains only to narrow pockets of the vocabulary, making it unclear whether a more general pressure towards arbitrariness may overpower such potential biases when considering a more semantically diverse selection of the vocabulary [2, 18].

A further issue with current studies of non-arbitrariness in sound-meaning correspondences is that, save for a single exception [19], cross-linguistic corpus studies of non-arbitrary associations have tended to rely on a small number of languages (maximally 200) and focusing on small semantically-restricted sets of words, ranging from phonation-related organs [20] to South American animals [14], to spatial orientation (demonstratives) [13, 21], repair initiators (like *huh?* in English) [22] and the conceptualization of magnitude in Australian languages [23]. These studies involve confirmatory analyses, aiming to test specific hypotheses regarding sound-meaning correspondences; as a consequence, they are guided by a priori intuitions or indirectly by findings from other disciplines. These limitations may help explain, at least in part, why language scientists typically consider non-arbitrary associations to be marginal phenomena that may only apply to small, strictly circumscribed regions of the vocabulary [3]. In this paper, we therefore conduct a comprehensive set of analyses involving a semantically diverse set of words from close to a two-thirds of the world's languages.

Testing associations on a global scale

The availability of a large collection of word lists allows us to search for statistically robust associations in an unsupervised, theory-neutral manner. The data consist of 28-40 lexical items from 6452 word lists, with a subset of 328 word lists having up to 100 items [24]. Words are transcribed into a phonologically simplified system consisting of 34 consonant and 7 vowels, which we refer to collectivelly as 'symbols' (see Table S1). These words belong to what is often referred to as 'basic vocabulary', including for instance pronouns, body part terms, property words, motion verbs and nouns describing natural phenomena [25]. The word lists include both languages and dialects, spanning 62% of the world's languages and about 85% of its lineages (see Fig. 1). A lineage is a maximal set of languages that can be shown to have a common ancestor. Such a set may have only one member (an isolate) or multiple members (a family).

Regarding the classification of languages, the Glottolog genealogical classification is preferable over other available alternatives because it is the only one to classify every living or extinct language while providing brief pointers to justifications for all choices taken—however, a less conservative independent classification was used additionally in the main test (see below). We stratify languages geographically by dividing the world's landmass into six largely independent linguistic macro-areas: North America, South America, Eurasia, Africa, Greater New Guinea and Australia—these regions have a history of attested contact within them but little contact between them in prehistorical times [26]. In order to guarantee that only truly global associations were selected, we screened the sound-meaning associations, keeping only those where the concept and symbol were attested in languages from at least 10 different lineages and found in no less than three different macro-areas.

We aim to capture robust and widespread tendencies in sound-meaning associations, where 'tendency' should be understood as a systematic bias in the frequency with which certain words tend to carry specific symbols in contrast to their baseline occurrence in other words. Crucially, a strong tendency does not imply that a signal has an extremely high frequency of occurrence, and conversely a very frequent sound-meaning co-occurrence is not sufficient evidence to discount chance. Importantly, whatever advantage a sound-meaning pairing might confer in terms of learning or processing, it has to be considered in the context of a myriad of competing factors that shape the phonetic and phonological fabric of words, from articulatory production costs [27] to systemic constraints due to the similarity with other lexical elements [28].

Our statistical approach consists in a series of tests where the presence of a symbol in a word is contrasted against a suitable subset of other words, and then the bias is evaluated across lineages. To begin, we calculate, for each concept and symbol, a genealogically balanced average ratio of the times they co-occur in a word of a language for which both symbol and concept are attested. We simulated the same quantity based on the rest of the concepts and compared it with the previously computed quantity (see Materials & Methods). The associated P-value roughly estimates the chance of finding the same or more extreme (genealogically balanced) average by picking any word other than the target one. Notice that this includes both recurring sound-meaning pairings as well as its complement, sound-meaning associations that are observed less often than expected given our null model.

Crucially, a sequence of tests need to be applied in order to ensure that potential associations are not statistical artifacts (see Materials & Methods and Supplemental Methods for more details). First, we used two independent worldwide language classifications with contrasting degrees of conservativeness [29, 30]. Second, we controlled the false discovery rate at a 5% expected level of false positives (for both classifications independently) so as to avoid an inflated number of associations due to multiple comparisons.

Third, word length is trivially correlated with the chance of finding any particular symbol. There is considerable variance in the (genealogically balanced) length of the words in our dataset, with some pronouns, negation and basic verbs (like *say* and *give*) consisting only of about three symbols on average, whereas the length of some color words and body part terms contain is over five (see Fig. S1). We filter out associations that also emerge when all of the symbols of all the words of each language are randomly permuted while keeping word lengths fixed.

Fourth, besides the mere number of symbols, word length might be a confound due to the fact that different phonotactic restrictions might apply accordingly. For instance, in a language that only allows CV structures and also prohibits the presence of word-initial liquids, no monosyllabic words will carry liquids. To remedy this, we performed a test similar to the first one described but this time comparing words only with the length-matched equivalents of different concepts.

Finally, in order to filter out associations due to areal contact or unresolved genealogy we looked for association that could be detected within the linguistic macro-areas independently. Thus, we restricted our attention to associations that passed all these statistical controls and for which a bias consistent with the world-wide trend could be found in at least three macro-areas, with no single area showing a bias in the opposite direction.

It should be noted that the overall testing scheme is conservative and that it is likely to have a large false negative rate. Also working against our analyses is the fact that the core set of concepts we use was originally gathered due to their exceptional phylogenetic persistence and resistance to borrowing, thus rendering them less likely to be adapted to potential func-

Table 1 Summary of signals found in the ASJP database. Positive and negative signals are those that have frequency significantly larger and smaller than expected.

Concept	Pos. symb	Neg. symb
ash	u	-
bite	k	-
bone	k	y
breasts	u m	a h r
dog	s	t
drink	-	a
ear	k	-
eye	-	a
fish	a	-
full	p b	-
hear	N	-
horn	k r	-
I	5	u p b t s r l
knee	o u p k q	-
leaf	b p l	-
name	i	o p
nose	u n	a
one	t n	-
red	r	-
round	r	-
sand	s	-
skin	-	m n
small	i C	-
star	z	-
stone	t	-
tongue	e E l	u k
tooth	-	b m
water	-	t
we	n	p l s
you	-	u o p t d q s r l

tional biases that might underlie specific sound-meaning associations. Moreover, it is not clear a priori whether the granularity of our phonetic descriptions is sufficiently fine to capture widespread sound-meaning relations—for instance, the opposition between voiced and unvoiced consonants and between rounded and unrounded in vowels have been suggested to bear importance for sound-symbolism [31, 21], but each feature pair are usually conflated under a single symbol in the database. For these reasons, the associations found in our analyses should be regarded as providing a lower-bound estimate of the presence of non-arbitrariness in sound-meaning pairings.

Strong worldwide associations

Our analysis detected 74 (positive and negative) sound-meaning associations, involving 30 concepts and 23 symbols. All of these associations are referred to as 'signals' (see Table 1; more detail is provided in Tables S2 and S3).

Signals will be described in terms of the most relevant information about them: the frequency of the symbol in the words corresponding to the concept (p), the ratio between that frequency and the frequency in other words (RR), the number of lineages that were analyzed for the global association (n_l) and the ratio between the number of areas where the association was independently found and the total number of tested areas (a_s/a_t).

Some concepts are associated with more than one signal. These are expected to be correlated; across languages it is often observed that there are preferences or restrictions with regard to the co-occurrence of symbols within one and the same word for either diachronic or synchronic phonotactic reasons. As an example, it is known that high front vowels trigger palatalization [32], so it is therefore not surprising that the voiceless palato-alveolar affricate C appears with i in the signals of *small*. In a set of testable pairs of signals (see Materials & Methods) signals sharing a concept tend to be significantly associated in about 41% of the time, against only 8% of signals involving different concepts (see Table S4)

The signals found in our analysis show a mixture of well-known and new associations. In line with the considerable literature on magnitude sound symbolism, the concept *small* was found to be associated with the high front vowel i (RR=1.58, p=.61, n_l=78, a_s/a_t=3/5), consistent with findings linking vowel height quality and size [13, 16], and with the palatal consonant C (RR=5.12, p=.41, n_l=61, a_s/a_t=3/4), also in agreement with previous work [13, 23].

We also observed a strong association between *round* and r-sounds (RR=2.48, p=.37, n_l=56, a_s/a_t=4/5). While most recent research has emphasized the role of consonants in shape-sound meaning associations like this [33, 34], the usual hypothesis in this direction concerned the correlation between vowel roundedness and round objects [10] – association that appears as a tendency in our analyses without reaching the minimum statistical threshold established before. Both *small* and *round* have been linked to the phenomenon of cross-modal mapping [9, 35, 12]. Another property word, *full*, is endowed with a pair of signals involving voiced (RR=1.91, p=.22, n_l=213, a_s/a_t=4/6) and unvoiced bilabial stops (RR=2.11, p=.13, n_l=231, a_s/a_t=5/6).

Some of the strongest signals found correspond to body parts. *Tongue* was very strongly associated with the lateral 'l' (RR=2.77, p=.41, n_l=280, a_s/a_t=6/6) and the mid and low front vowels e (RR=1.54, p=.11, n_l=322, a_s/a_t=5/6) and E (RR =1.73,p=.11 ,n_l=164, a_s/a_t=4/6). *Nose* was found to be associated most strongly with the alveolar nasal n (RR=1.47, p=.35, n_l=334, a_s/a_t=4/6), the high back vowel u (RR=1.38, p=.35, n_l=325, a_s/a_t=4/6). The link between *nose* and nasality has been noted previously [36], in particular in reference to the conjecture that body part terms used in phonation makes use of the distinctive qualities provided by the relevant organ [20].

Breasts was associated with the bilabial nasal consonant m (RR=1.63, p=.32, n_l=320, a_s/a_t=4/6) and the high back

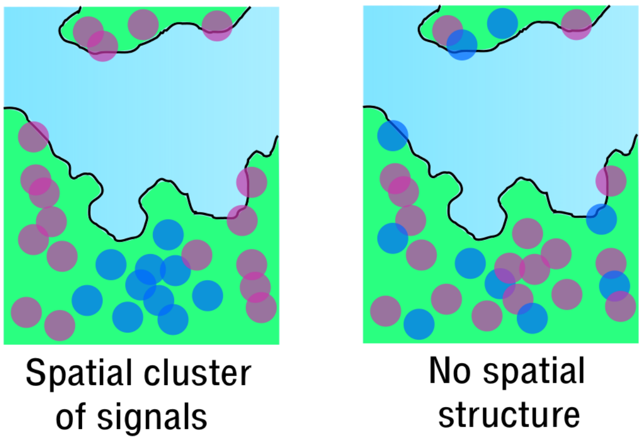


Fig. 2. Competing configurations of the spatial distribution of the tested languages. Blue and fuchsia dots represent languages with and without a specific signal, respectively. In the panel to the left, the likelihood of a language having the signal is correlated with its geographical distance to its nearest neighbor, and on the right there is no spatial structure.

vowel u (RR=1.46, p=.37, n_l=317, a_s/a_t=4/6). Similar associations were found in the nursery terms for *mother*, a concept with which it often colexifies. It has been suggested that this might be due to the mouth configuration of suckling babies or to the sounds feeding babies produce [37, 38].

While this study lends support to a number of associations that were either elicited in experiments or conjectured based on a much smaller number of languages, it also provides telling negative evidence on others. Together with the association between high front vowels and the concept of small, there has been reports on a connection between back low vowels and the notion of big [21]. However, *big* (n_l=73) and *large* (n_l=74) and o did not show any relevant signature of association in our sample at the global level. Similarly, an analogous front/back vowel opposition has been proposed to hold between proximal and distal pronouns—the purported explanation being that proximal referents tend to be small whereas distal referents are usually large [21]. The concepts *this* (n_l=71) and *that* (n_l=74), however, do not show any associations with i and o (respectively).

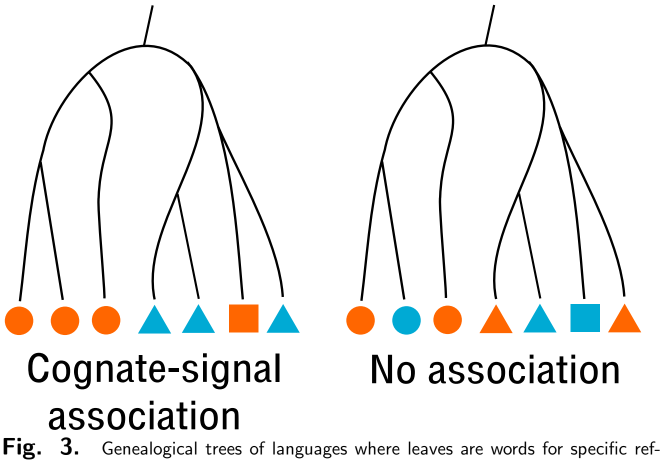


Fig. 3. Genealogical trees of languages where leaves are words for specific referents. In the figure to the left, cognate classes (depicted as different shapes) are associated with signal presence (blue shapes), whereas to the right there is no such correspondence.

Origins and nature of the associations

As discussed in the previous sections, there are multiple theories which attempt to elucidate why humans find that some sounds are more convenient or salient in association with certain meanings. How these hypothesized mechanisms lead to the widespread biases in vocabularies we find here is a complex question that is unlikely to be fully answered by the inspection of wordlists. Nonetheless, we can attempt to evaluate some of the potential consequences of those theories given the coarse level of detail of our data.

Functional advantages might increase the likelihood of signals being borrowed across languages in contact with one another, thus producing spatial diffusion patterns [38] (see Figure 2). The existence of opposing factors obscure definitive inferences in this direction, though: basic vocabulary items are particularly resistant to borrowing but unresolved genealogy involving nearby languages would be confounded with borrowing. In the same direction, large populations have been claimed to be more efficient at gaining and retaining non-arbitrary sound-meaning associations given a potential functional value [38], which is coherent with recent evidence from some Austronesian languages showing that larger populations gain new words at a faster rate [39].

We determined whether present-day log population size and log distance to the nearest genealogically unrelated language bearing the (positive) signal are effective predictors for signal presence, via a mixed effects logistic model (see Table S6 and Supplemental Methods). At $\alpha = 0.05$, log population turned out to be significant in about one third of the cases, but the effect was small and as many times positive as it was negative, which rules out a consistent role for population. Only one fifth of the signals showed sensitivity to the distance of nearest neighbors with signal, with all of the cases having an effect in the predicted direction by our model. On average, and in contrast to the case in which a language and its signal-bearing nearest genealogically unrelated neighbor are spoken in exactly the same place, the probability of finding the signal also in the language drops by 28%.

From a historical perspective, it has been suggested that sound-meaning associations might be evolutionarily preserved features of spoken language [40], potentially hindering regular sound change [16]. Furthermore, it has been claimed that widespread sound-meaning associations might be vestiges of one or more large-scale prehistoric proto-languages [15]. Tellingly, some of the signals found here feature prominently in reconstructed "global etymologies" [41, 42] that have been used for deep phylogeny inference [43]. If signals are inherited from an ancestral language spoken in remote prehistory, we might expect them to be distributed similarly to inherited, cognate words; that is, their distribution should to a large extent be congruent with the nodes defining their linguistic phylogeny (see Figure 3 for illustration).

A direct evaluation of this hypothesis is infeasible due to the absence of etymological dictionaries for all but a few families. However, it can be tested indirectly given that cognate words are expected to be more similar to one another than non-cognates [44]. We investigated whether the presence of the signal-bearing symbol was a better indicator of overall form similarity between words than other shared symbols, using a beta mixed regression model that distinguishes the effects of symbols, concept and lineage (see Supplemental Methods). The model is heavily dominated by the effect of lineage, and signal presence (while significant) has a negligible effect in the opposite direction than predicted: the genealogically balanced average effect is less than a 0.5% decrease in similarity for those words sharing a signal-related symbol compared to those sharing some other symbol.

Consistency in word position is important for establishing cognacy [44, 45]. Further support for the idea that signals are not residuals of deep history comes from the analysis of the position within the word in which they occur, in particular whether they have a clear word-initial bias. All in all, we find that signals do not have a consistent cross-linguistic preference or dispreference in this respect beyond well-established cross-linguistic phonotactic patterns, such as the avoidance of liquids or the prevalence of dorsal and labial stops in word-initial position [46, 47] (see Supplemental Methods and Table S5).

These results suggest that although it is possible that the presence of signals in some families are symptomatic of a particularly pervasive cognate set, this is not the usual case. Hence, the explanation for the observed prevalence of sound-meaning associations across the world has to be found elsewhere [48].

Conclusion

We have demonstrated that a substantial proportion of words in the basic vocabulary are biased to carry or to avoid specific sound segments, both across continents and linguistic lineages. Given that our analyses suggest that phylogenetic persistence or areal dispersal are unlikely to explain the widespread presence of these signals, we are left with the alternative that the signals are due to factors common to our species, such as sound symbolism, iconicity, communicative pressures or synaesthesia. We expect future research to further elucidate the role and interaction of these factors in driving the observed sound-meaning association biases, and to extend the scope of our findings to a broader portion of the vocabulary.

The outcome of our analyses have consequences for historical-comparative linguistics, where it has been suggested that there is a small set of ultra-conserved words that are particularly useful for establishing ancient genealogical relations beyond the limits of the comparative method [43]. However, some of these words are involved in the signals discovered here: *we* is associated with the alveolar nasal, *hear* with the velar nasal, and *ash* with the vowel u. Thus, proposals of far-reaching etymologies based on words of similar form and meaning should be accompanied by an evaluation of whether the observed lexical similarities might have resulted from the kinds of signal discussed in this paper rather than common inheritance. More generally, even though it is unclear whether the locus of the emergence of signals is in the invention or historical development of lexical roots, our findings have implications for the study of the dynamics of lexical phonology.

In summary, our results provide new insights into the constraints that affect how we communicate, suggesting that despite the immense flexibility of the worlds languages, some sound-meaning associations are preferred by culturally, historically and geographically diverse human groups.

Materials and Methods

Basic vocabulary word lists. The data set used for this study is drawn from version 16 of the ASJP database [24]. ASJP comprises 6895 word lists from around 62% of the world's languages, covering 85% of families, isolates, and unclassified languages (using the Ethnologue [51] for these statistics). After removing artificial languages, pidgins and creoles and varieties whose ISO-639-3 code cannot be confirmed, the number goes down to 6447 word lists, corresponding to 4298 different languages and 359 lineages. The database was not constructed for the specific purpose of studying sound symbolism, but rather for identifying genealogical relations among languages. For this reason, it generally consists of the 40-item subset of the 100-item so-called Swadesh list [52] that are assumed to remain stable as languages diverge into different lineages over time [53]. Of these word lists, 328 additionally contain the remaining 60 Swadesh lists items.

Words are rendered in a unified transcription system, which facilitates cross-linguistic comparison but also ignores phonetic details such as vowel length, nasalization, tones, and retroflexation. Vowel quality distinctions are merged into seven categories (high front, mid front, low front, high-mid central, low central, high back, mid-low back) (see [54] for a discussion of the system).

Each 40-item word list provides translational equivalents, when available, for the following items: *blood, bone, breast, come, die, dog, drink, ear, eye, fire, fish, full, hand, hear, horn, I, knee, leaf, liver, louse, mountain, name, new, night, nose, one, path, person, see, skin, star, stone, sun, tongue, tooth, tree, two, water, we, you (sg).* The additional Swadesh list items contained in some of the word lists are: *all, ash, bark, belly, big, bird, bite, black, burn, claw, cloud, cold, dry, earth, eat, egg, feather, flesh, fly, foot, give, good, grease, green, hair, head, heart, hot, kill, know, lie, long, man, many, moon, mouth, neck, not, rain, red, root, round, sand, say, seed, sit, sleep, small, smoke, stand, swim, tail, that, this, walk, what, white, who, woman, yellow.*

Associations between symbols and concepts. The fundamental statistic in our analysis is p_{ij}, the maximum likelihood estimator (i.e. the sample frequency) for the probability of finding that concept i has at least one instance of symbol j, after randomly choosing a lineage, a language within the lineage and a dialect within the language (if any) in that sequential order. Naturally, this calculation is restricted to the set of dialects of languages for which the concept and the phone are attested (which we will refer as S_{ij}); for each of those sets this quantity is formally:

$$p_{ij} = \frac{1}{|L|}\sum_{k=1}^{|L|}\left(\frac{1}{|L_k|}\sum_{l=1}^{|L_k|}\frac{1}{|L_{kl}|}\sum_{d=1}^{|L_{kl}|}\pi_{ij}^{kld}\right)$$

The sets L, L_k and L_{kl} are the sets of all lineages, languages within lineage k and dialects of language l within lineage k. π_{ij}^{kld} is a binary variable that takes value 1 if there is at least one instance of symbol j in the word for concept i for dialect d of language l from lineage k (always within the set S_{ij}) and 0 otherwise.

This computation is conservative in that all languages known to belong to the same genealogical group influence the aggregated statistics in the same way regardless of their size, but on the other hand it guarantees the minimum possible bias in the dependence of the languages' words. In order to avoid testing cases whose coverage is insufficiently wide before testing, we evaluated only those associations for which S_{ij} comprises ten lineages in each of three different macro-areas at least.

Conversely, for each dialect of each language we calculated the proportion of words other than that associated with i that have symbol j, and we note this as π_{-ij}^{kld}, and similarly the genealogical balanced average as p_{-ij}. These probabilities are used to produce $n_{sim} = 1000$ Monte Carlo simulations of symbol j presence/absence for all the languages in S_{ij} - the set of p_{-ij} values resulting from these simulations will be called ζ_{ij}. The purpose is to compare ζ_{ij} with π_{ij} in order to answer the question: does symbol j appear much more (or much less) often when a subset of words referring to concept i is selected than in a randomly picked set of words from the same languages? The two-tailed P-value for a particular concept i and symbol j is then [55]

$$P = \frac{1}{n_{\mathrm{sim}}+1}\left(2\min\{|x\in\zeta_{ij} : x \geq p_{ij}|, |x\in\zeta_{ij} : x\leq p_{ij}|\}+1\right)$$

where $|\cdot|$ is the cardinality of the set.

The large number of tests performed require a control for type I errors. We perform a False Discovery Rate (FDR) analysis fixing the FDR rejection threshold to .05, which means that we will allow no more than 5% of false positives on average. For this purpose we use the method described in [56]. The basic idea is that the distribution of P-values comes from a mixture of a uniform distribution (that corresponds to the baseline of tests where no associations beyond chance are present) and a distribution concentrated near $P = 0$ of true positives. The method used here learns the mixture proportion of the uniform distribution from values P from 1 down to a threshold that is adjusted in order to reduce the false non-discovery rate (FNDR).

This entire procedure was repeated with a different, less conservative, genealogical classification—the one provided by the World Atlas of Language Structures (WALS) [29]. For our analysis, we only considered associations that were below the defined FDR level according to both classifications. The fraction of the component of true negatives learned from both classifications was around 0.65.

Regarding possible confounds due to word length, we performed two extra tests on those associations that successfully passed the previous test. First, we repeated the same global test using the Glottolog classification this time comparing p_{ij} with simulations obtained from words of exactly the same number of symbols in each language (and dialect). Second, for each language (and dialect) in S_{ij}, $n = 1000$ of independent simulations we sampled without replacement as many random symbols from words other than i up to the length of word i. This effectively produces, for each word i, a random counterpart equivalent to shuffling all the symbols corresponding to all the the words of a language while keeping word lengths constant. Over each of those sets, the same association test based on the Glottolog classification was performed. In both of these procedures, we imposed a stricter cutoff: if any of the simulations yield a value of p_{ij} equally or more extreme, we would reject the association as of potential interest.

Finally, for each macro-area with at least 10 independent lineages in S_{ij}, we analyzed the presence of a significant direction of association as in the main associations test—computing both empirical and random probabilities using only the languages of that area—with the difference that we flagged each macro-area specific association with $P \leq .1$. It should be noticed that this does not imply a softer rejection threshold than in the worldwide case: we only keep associations that display a bias consistent with the world-wide trend in at least half of the macro-areas, with the extra condition that no macro-area should exhibit a bias in the opposite direction.

To summarize: only associations that successfully satisfied all the requirements of the overall association test (with Glottolog and WALS classifications independently), the word length and the matched-length tests, and for which a consistent preference in at least half of the macro-areas could be found were considered "signals".

Association between signals. As in the previous case, we analyze sets of languages for which both the concept and the symbol associated with a pair of signals was present in at least ten lineages in each of (at least) three macro-areas. The association between signals—which we will refer to A and B here—was tested by means of a simple mixed effects logistic model,

$$\mathrm{logit}(\text{signal A presence}) = \alpha_{\text{signal B presence}} + \alpha^{\text{lineage}}$$

where $\alpha_{\text{signal A presence}}$ is the coefficient related to the presence of signal A, and α^{lineage} is a random coefficient structured according to lineage. To the results obtained by comparing all the pairwise associations between signals belonging to the core 40 words, we applied a threshold on the FDR of 5%. About 12% of the 2062 cases satisfied this condition. The results of associations regarding same-concept signals and the genealogically balanced average effect on the presence of signal B on A can be found in Table S4.

ACKNOWLEDGMENTS. We acknowledge the comments of Bernard Comrie, Brent O. Berlin, Stephen C. Levinson, Mark Dingemanse, Russell Gray and Eric W. Holman. We also thank Jeremy Collins and Stefany Moreno for assistance with other aspects of the manuscript. HH's research was made possible thanks to the financial support of the Language and Cognition Department at the Max Planck Institute for Psycholinguistics, Max-Planck Gesellschaft, and a European Research Council's Advanced Grant (269484 "INTERACT") to Stephen C. Levinson. SW's research was supported by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC Grant Agreement No. 295918 and by a subsidy of the Russian Government to support the Program of Competitive Development of Kazan Federal University.

1. Saussure, FD (1916) Cours de linguistique générale, ed Bally C, Sechehaye A, Riedlienger A (Payot, Paris).
2. Hockett CF (1960) The origin of speech. Sci Am 203:89–96.
3. Pinker S (1999) Words and Rules: The Ingredients of Language. (Perseus Books, New York).
4. Dingemanse M (2012) Advances in the cross-linguistic study of ideophones. Language Linguist Compass 6:654–672.
5. Dingemanse M, Blasi DE, Lupyan G, Christiansen MH, Monaghan P (2015). Arbitrariness, iconicity, and systematicity in language. Trends Cogn Sci 19(10):603–615.
6. Bergen BK (2004) The psychological reality of phonaesthemes. Language 80:290–311.
7. Monaghan P, Christiansen MH, Chater N (2007) The phonological-distributional coherence hypothesis: cross-linguistic evidence in language acquisition. Cognitive Psychol 55:259–305.
8. Köhler, W (1929) Gestalt Psychology (Liveright, New York).
9. Ramachandran VS, Hubbard EM (2001) Synaesthesia—a window into perception, thought and language. J Consciousness Stud 8:3–34.
10. Maurer D, Pathman T, Mondloch CJ (2006) The shape of boubas: soundshape correspondences in toddlers and adults. Developmental Sci 9:316–322.
11. Ludwig VU, Adachi I, Matsuzawa T (2011) Visuoauditory mappings between high luminance and high pitch are shared by chimpanzees (Pan troglodytes) and humans. Proc Natl Acad Sci USA 108(51):20661-20665.
12. Cuskley C, Kirby S (2013) Synaesthesia, cross-modality and language evolution. Oxford Handbook of Synaesthesia, ed Simner J, Hubbard EM (Oxford University Press, Oxford, UK), pp 869–907.
13. Hinton L, Nichols J, Ohala JJ (eds.) (2006) Sound Symbolism. (Cambridge University Press, New York).
14. Berlin B (2006) The first congress of ethnozoological nomenclature. J Roy Anthropol Inst 12(s1):S23–S44.
15. Imai M, Kita S (2014) The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. Philos T Roy Soc B 369(1651):20130298.
16. Nuckolls JB (1999) The case for sound symbolism. Annu Rev Anthropol 28:225–252.
17. Reilly J, Hung J, Westbury C. (2016) Non arbitrariness in mapping word form to meaning: crosslinguistic formal markers of word concreteness. Cognitive Sci. doi: 10.1111/cogs.12361.
18. Monaghan P, Christiansen MH, Fitneva SA (2011) The arbitrariness of the sign: learning advantages from the structure of the vocabulary. J Exp Psychol Gen 140:325–347.
19. Wichmann S, Holman EW, Brown CH (2010) Sound symbolism in basic vocabulary. Entropy 12:844–858.
20. Urban M (2011) Conventional sound symbolism in terms for organs of speech: a cross-linguistic study. Folia Linguist 45:199–214.
21. Johansson N, Zlatev J (2013) Motivations for sound symbolism in spatial deixis: a typological study of 101 languages. The Public J Sem 5:3–2.
22. Dingemanse M, Torreira F, Enfield, NJ (2013). Is "Huh?" a universal word? Conversational infrastructure and the convergent evolution of linguistic items. PLOS ONE, 8(11), e78273.
23. Haynie H, Bowern C, LaPalombara H (2014) Sound symbolism in the languages of Australia. PLOS ONE 9, e92852.
24. Wichmann S et al. (2013) The ASJP database (version 16). http://asjp.clld.org/. Accessed on 2013-07-02.
25. Tadmor U, Haspelmath M, Taylor B (2010) Borrowability and the notion of basic vocabulary. Diachronica 27:226–246.
26. Hammarström H, Donohue M (2014) Some principles on the use of macro-areas in typological comparison. Lang Dynam Change 4:167-187.
27. Napoli D J, Sanders N, Wright R (2014). On the linguistic effects of articulatory ease, with a focus on sign languages. Language, 90(2), 424-456.
28. Vitevitch MS, Luce PA (2016). Phonological neighborhood effects in spoken word perception and production. Annu Rev Linguistics 2:75–94.
29. Haspelmath M, Dryer MS, Gil D, Comrie B (2005) The World Atlas of Language Structures (Oxford University Press, Oxford, UK).
30. Nordhoff S, Hammarström H, Forkel R, Haspelmath M (2013) Glottolog 2.1, http://glottolog.org. Accessed on 2013-07-02.
31. Lockwood G, Dingemanse M (2015) Iconicity in the lab: a review of behavioral, developmental, and neuroimaging research into sound-symbolism. Front Psychol 6. http://dx.doi.org/10.3389/fpsyg.2015.01246
32. Bateman N (2011) On the typology of palatalization. Lang Linguist Compass 5:588–602.
33. Nielsen A, Rendall D (2011) The sound of round: evaluating the sound-symbolic role of consonants in the classic Takete-Maluma phenomenon. Can J Exp Psychol 65:115–124.
34. Fort M, Martin A, Peperkamp S (2015). Consonants are more important than vowels in the Bouba-kiki effect. Lang Speech 58(2):247–266.
35. Bankieris K, Simner J (2015) What is the link between synaesthesia and sound symbolism? Cognition 136:186–195.
36. Greenberg JH, Ferguson CA, Moravcsik EA (Eds.). (1978) Universals of human language: phonology (Vol. 2). (Stanford University Press, Stanford, CA).
37. Jakobson R (1960) Why "mama" and "papa". Perspectives in Psychological Theory, Dedicated to Heinz Werner, eds Kaplan B, Wapner S (International Universities Press, New York), pp 124–134.
38. Traunmüller H (1994) Sound symbolism in deictic words. Tongues and Texts Unlimited. Studies in Honour of Tore Jansson on the Occasion of his Sixtieth Anniversary, eds. Janson T, Aili H, af Trampe P (Stockholms Universitet, Institutionen för klassiska språk, Stockholm), pp 213-234.
39. Bromham L, Hua X, Fitzpatrick TG, Greenhill SJ (2015) Rate of language evolution is affected by population size. P Natl Acad Sci USA 112(7):2097–2102.
40. Nygaard LC, Cook AE, Namy LL (2009) Sound to meaning correspondences facilitate word learning. Cognition 112:181–186.
41. Ruhlen M (1994) On the Origin of Languages: Studies in Linguistic Taxonomy (Stanford University Press, Stanford, CA).
42. Starostin SA, Bronnikov Y (19982009) Languages of the World Etymological Database. Available at http://starling.rinet.ru/cgi-bin/main.cgi?flags=eygtnnl, Part of the Tower of Babel Evolution of Human Language Project.
43. Pagel M., Atkinson QD, Calude AS, Meade A (2013) Ultraconserved words point to deep language ancestry across Eurasia. Proc Natl Acad Sci USA 110:8471–8476.
44. Steiner L, Stadler PF, Cysouw M (2011) A pipeline for computational historical linguistics. Lang Dynam Change 1:89–127.
45. Jäger, G (2015) Support for linguistic macrofamilies from weighted sequence alignment. P Natl Acad Sci USA 112(41):12752–12757.
46. Proctor MI (1995) Gestural characterization of a phonological class: The liquids. PhD dissertation, Yale University.
47. MacNeilage PF, Davis BL (2000) On the origin of internal structure of word forms. Science 288(5465):527–531.
48. Campbell L, Poser WJ (2008) Language Classification (Cambridge University Press, New York).
49. Monaghan P, Shillcock RC, Christiansen MH, Kirby S (2014) How arbitrary is language? Philos T R Soc B 369:20130299.
50. Perry LK, 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.
51. Lewis P, Simons G, Fennig C (2013) Ethnologue: Languages of the World (SIL International, Dallas, TX).
52. Swadesh M (1955) Towards greater accuracy in lexicostatistic dating. Int J Am Linguist 21:121–137.
53. Holman EW, Wichmann S, Brown CH, Velupillai V, Müller V, Bakker D (2008) Explorations in automated language classification. Folia Linguist 42:331–354.
54. Brown CH, Holman EW, Wichmann S (2013) Sound correspondences in the world's languages. Language 89:4–29.
55. North BV, Curtis D, Sham PC (2002) A note on the calculation of empirical P values from Monte Carlo procedures. Am J Hum Gen 71:439–441.
56. Strimmer, K (2008) fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. Bioinformatics 24(12):1461–1462.

Supplemental Methods

Positional test. We simulate, for each language and signal, random positions of the relevant signal-associated symbol based on all the available positions in the word according to the consonant/vowel distinction. Concretely, we calculate the number of times the phone is initial when its simulated counterpart is not, averaging genealogically and respecting the vowel and consonant template of each word. Then we compare this quantity in the original word list against $n = 1000$ simulations and consider those cases in which the original bias is larger than 95% of the simulated cases. These results can be observed in Table S5.

Areal and population test. For each positive signal we calculated the great circle distances—i.e., the distance in kilometers of the shortest geodesic connecting two points in the surface of the Earth—involving all languages having both the relevant symbol and concept (but not necessarily the signal) and their nearest language from a different lineage that has the (positive) signal (dnn). The hypothesis is that small distance from a language that has a signal will influence the likelihood of signal presence in a given language. Only signals belonging to the group of 28-40 better attested concepts were used for the analysis, and only one dialect per language was chosen. Extinct languages were excluded from the analyses.

For the testing we used a generalized logistic model with random effects:

$$\begin{aligned}\operatorname{logit}(\mathbb{E}[\text{signal presence}]) = \alpha + (\beta_{\text{dnn}} + \beta_{\text{dnn}}^{\text{lineage}})\log(1+\text{dnn})\\ + \beta_{\text{pop}}\log(\text{population}) + \alpha^{\text{lineage}}\end{aligned}$$

where the superscripted coefficients ($\beta_{\text{dnn}}^{\text{lineage}}$ and α^{lineage}) are random effects structured according to the lineage. Lineage as a random intercept is introduced as a means of accounting for the varying baseline presence of the signals within lineages, and their presence as random slopes aims to capture the fact that lineages have spread with different rates across the globe. The logarithmic transforms aims to reduce the effect of population and distance outliers. P-values were estimated through an asymptotic likelihood ratio test. Apart from the estimated coefficients, we calculated the genealogical balanced mean difference in probability of having a signal for two reference points, one variable at a time. For population, the difference was calculated between fixing all languages' populations to 10000 individuals and a single individual, and for dnn between 1000 km—which is roughly the maximum radius of linguistic areas as defined in AUTOTYP [1]—and 0 km (which correspond to the situation where both languages as spoken at the same place). The results can be observed in Table S6.

Word similarity test. Ideally, a proper phylogenetic test in the context of language history would comprise some kind of data carrying a phylogenetic signal (like cognate sets or collections of regular sound changes) and a sound evolutionary model that would lead to a tree or a distribution of trees. Unfortunately, such trees exist for only a handful of language families [2,3]. Instead, we approach the question of both phylogenetic stability and ancestry of signals by analyzing word form similarity, which serves as a proxy for cognacy. If it is a correct hypothesis that signals render words less prone to change and that they are prehistoric vestiges, then, after controlling for concept, symbol, and lineage, we would expect to find that the similarity among words is predicted by signals.

The distance between words used here is the Levenshtein distance, which has found several uses in linguistics and often correlates with perceptual, processing and other meaningful lexical distances differences [4,5]. The Levenshtein distance between strings x and y LD(x,y) is defined as the minimum number of edits, additions or deletions of characters necessary to make two strings identical. For instance, 'Zultus' and 'sulus'—*star* in Uyghur and Sakha (two Turkic languages) respectively, have a Levenshtein distance of 2: a change of 'Z' to 's' and the deletion of 't' in the Sakha word. The normalized Levenshtein distance is simply $l = LD(x,y)/\max(|x|,|y|)$

For every family with at least six languages and every combination of concept and symbol, we calculated the Levenshtein distance between all members of two groups: word pairs for a concept belonging to a combination, and word pairs for a concept sharing at least one symbol but not the symbol relevant for the combination. For instance, given a family with three languages having the forms *ana*,*ena* and *ete* for the concept "rock", and considering the combination rock-n, we will have the two following groups: (a**n**a,e**n**a) and (**e**na,**e**t**e**). Families with less than three distances in any of the groups were excluded from the analysis.

In order to summarize the previous information, we calculated, for each family, the probability of choosing a distance in the signal-sharing group and another in the non-signal-sharing group and finding that the first is smaller than the second ($\Pr(l_s < l_{-s})$). The larger this quantity, the more reliable an estimator of wordform similarity the association is.

Then we implemented the following beta regression mixed model with logistic link function and constant precision parameter:

$$\begin{aligned}\operatorname{logit}(\mathbb{E}[\Pr(l_s < l_{-s})]) = \sum_{\text{concepts}} \beta_i I_i + \sum_{\text{symbols}} \beta_j I_j\\ + \alpha_{\text{signalhood}} + \alpha^{\text{lineage}}\end{aligned}$$

where the i and j indexes run over the set of concepts and symbols, respectively, the coefficient "signalhood" indicates whether the combination of concept and symbol is to be found in Table S2. 'signalhood' was coded as a single level common to all individual positive signals. α^{lineage} stands for a random intercept according to lineage. In order to cope with a few values of $\Pr(l_s < l_{-s})$ identical to 1 (that account for less than 0.5% of the data) we applied the transformation $t(x) = (x(N-1)+0.5)/N$ to the values [6]. As a way of accounting for the more robust evidence provided by lineages with a large number of distance pairs to be compared, we included a weight for each observation equal to the logarithm of the number of such pairs involved—however, the results did not differ considerably from the unweighted case. Overall, the model quality is heavily dominated by lineage: 86% vs. 3% of explained deviance with and without the lineage random effect, respectively.

References

[1] Nichols J, Witzlack-Makarevich A, Bickel B (2013) The AUTOTYP genealogy and geography database: 2013 release. (University of Zürich, Zürich). Available online: http://www.autotyp.uzh.ch/

[2] Dunn M, Greenhill SJ, Levinson SC, Gray RD. (2011) Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473:79–82.

[3] Gray RD, Drummond AJ, Greenhill SJ. (2009) Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323:479–483.

[4] Gooskens C, Heeringa W (2004) Perceptive evaluation of Levenshtein dialect distance easurements using Norwegian dialect data. *Lang Var Change* 16:189–207.

[5] Nerbonne J, Heeringa W, Kleiweg P (1999) Edit distance and dialect proximity. *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, eds Sankoff D, Kruskal J (CSLI Press, Stanford, CA), pp v–xv.

[6] Smithson M, Verkuilen J (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychol Methods* 11(1):54–71.

Supplemental Table 1. ASJP symbols and their description. IPA equivalents of the symbols can be found in Tables 1-2 of [54].

Symbol	Description
p	voiceless bilabial stop and fricative
b	voiced labial stop and fricatve
m	bilabial nasal
f	voiceless labiodental fricative
v	voiced labiodental fricative
8	voiceless and voiced dental fricative
4	dental nasal
t	voiceless alveolar stop
d	voiced alveolar stop
s	voiceless alveolar fricative
z	voiced alveolar fricative
c	voiceless and voiced alveolar fricative
n	voiceless and voiced alveolar nasal
S	voiceless postalveolar fricative
Z	voiced postalveolar fricative
C	voiceless palato-alveolar affricative
j	voiced palato-alveolar affricate
T	voiceless and voiced palatal stop
5	palatal nasal
k	voiceless velar stop
g	voiced velar stop
x	voiceless and voiced velar fricative
N	velar nasal
q	voiceless and voiced uvular stop
X	voiceless and voiced uvular fricative, voiceless and voiced pharyngeal fricative
7	voiceless glottal stop
h	voiceless and voiced glottal fricative
l	voiced alveolar lateral approximate
L	all other laterals
w	voiced bilabial-velar approximant
y	palatal approximant
r	all varieties of "r-sounds"
i	high front vowel, rounded and unrounded
e	mid front vowel, rounded and unrounded
E	low front vowel, rounded and unrounded
3	high and mid central vowel, rounded and unrounded
a	low central vowel, unrounded
u	high back vowel, rounded and unrounded
o	mid and low back vowel, rounded and unrounded

Supplemental Table 2. Complete list of positive signals found in the ASJP database. The column ‘Areal ratio’ indicates the ratio between the number of areas where the signals are independently found with respect the total number of areas with minimum coverage. RR stands for “risk ratio”. Family counts come from Glottolog [30].

Concept	Symb.	p_{ij}	p_{-ij}	$\sigma(p_{-ij})$	Δ	RR	Lineages	Areal ratio
ash	u	0.516	0.270	0.043	0.25	1.91	68	3/5
bite	k	0.438	0.259	0.042	0.18	1.69	73	3/5
bone	k	0.311	0.223	0.016	0.09	1.39	333	3/6
breasts	u	0.376	0.257	0.018	0.12	1.46	317	4/6
breasts	m	0.326	0.200	0.016	0.13	1.63	320	4/6
dog	s	0.225	0.128	0.015	0.10	1.76	285	3/5
ear	k	0.319	0.224	0.017	0.09	1.42	338	4/6
fish	a	0.613	0.524	0.019	0.09	1.17	327	3/6
full	p	0.255	0.121	0.016	0.13	2.11	231	5/6
full	b	0.229	0.120	0.016	0.11	1.91	213	4/6
hear	N	0.199	0.127	0.018	0.07	1.57	182	3/6
horn	k	0.339	0.222	0.019	0.12	1.53	221	4/6
horn	r	0.271	0.155	0.019	0.12	1.75	191	3/6
I	5	0.129	0.063	0.015	0.07	2.06	136	4/6
knee	u	0.472	0.256	0.018	0.22	1.84	303	4/6
knee	o	0.406	0.239	0.017	0.17	1.70	291	4/6
knee	p	0.218	0.121	0.014	0.10	1.81	278	5/6
knee	k	0.374	0.226	0.018	0.15	1.66	305	5/6
knee	q	0.313	0.136	0.027	0.18	2.30	73	3/5
leaf	p	0.232	0.119	0.014	0.11	1.94	290	3/6
leaf	b	0.185	0.124	0.014	0.06	1.48	274	3/6
leaf	l	0.268	0.154	0.016	0.11	1.75	270	4/6
name	i	0.474	0.378	0.020	0.10	1.25	320	3/6
nose	u	0.351	0.255	0.018	0.10	1.38	325	4/6
nose	n	0.356	0.242	0.016	0.11	1.47	334	4/6
one	t	0.266	0.178	0.015	0.09	1.49	343	3/6
one	n	0.320	0.248	0.017	0.07	1.29	348	3/6
red	r	0.350	0.156	0.037	0.19	2.24	61	3/5
round	r	0.371	0.149	0.038	0.22	2.48	56	4/5
sand	s	0.325	0.126	0.034	0.20	2.58	65	3/5
small	i	0.613	0.389	0.043	0.22	1.58	78	3/5
small	C	0.416	0.081	0.029	0.33	5.12	61	3/4
star	z	0.158	0.063	0.018	0.10	2.52	96	3/5
stone	t	0.239	0.181	0.015	0.06	1.32	340	3/6
tongue	e	0.339	0.220	0.017	0.12	1.54	322	5/6
tongue	E	0.278	0.161	0.020	0.12	1.73	164	4/6
tongue	l	0.419	0.151	0.017	0.27	2.77	280	6/6
we	n	0.380	0.246	0.017	0.13	1.54	325	3/6

Supplemental Table 3. Complete list of negative signals found in the ASJP database. The column ‘Areal ratio’ indicates the ratio between the number of areas where the signals are independently found with respect the total number of areas with minimum coverage. RR stands for “risk ratio”. Family counts come from Glottolog [30]

Concept	Symb.	p_{ij}	p_{-ij}	$\sigma(p_{-ij})$	Δ	RR	Lineages	Areal ratio
bone	y	0.065	0.122	0.013	-0.06	0.54	312	3/6
breasts	a	0.422	0.524	0.020	-0.10	0.81	329	3/6
breasts	h	0.093	0.149	0.016	-0.06	0.62	254	3/6
breasts	r	0.083	0.175	0.015	-0.09	0.47	290	3/6
dog	t	0.106	0.182	0.015	-0.08	0.58	337	4/6
drink	a	0.421	0.533	0.020	-0.11	0.79	310	4/6
eye	a	0.423	0.527	0.018	-0.10	0.80	357	4/6
I	u	0.116	0.262	0.018	-0.15	0.44	328	5/6
I	p	0.021	0.122	0.014	-0.10	0.18	297	5/6
I	b	0.030	0.124	0.014	-0.09	0.24	276	4/6
I	t	0.079	0.181	0.016	-0.10	0.44	332	4/6
I	s	0.036	0.131	0.015	-0.10	0.27	279	4/5
I	l	0.030	0.161	0.016	-0.13	0.19	277	6/6
I	r	0.061	0.177	0.015	-0.12	0.35	294	6/6
name	o	0.169	0.254	0.018	-0.09	0.67	297	4/6
name	p	0.049	0.122	0.015	-0.07	0.40	283	3/6
nose	a	0.391	0.524	0.019	-0.13	0.75	339	4/6
skin	m	0.109	0.207	0.016	-0.10	0.53	323	4/6
skin	n	0.170	0.256	0.016	-0.09	0.66	329	4/6
tongue	u	0.164	0.264	0.017	-0.10	0.62	327	3/6
tongue	k	0.167	0.232	0.017	-0.07	0.72	334	4/6
tooth	b	0.054	0.126	0.014	-0.07	0.43	282	4/6
tooth	m	0.130	0.205	0.016	-0.08	0.63	335	4/6
water	t	0.066	0.184	0.015	-0.12	0.36	345	6/6
we	p	0.052	0.121	0.015	-0.07	0.43	288	5/6
we	l	0.064	0.160	0.016	-0.10	0.40	268	5/6
we	s	0.077	0.129	0.015	-0.05	0.60	273	3/5
you	u	0.149	0.259	0.017	-0.11	0.58	316	3/6
you	o	0.165	0.246	0.017	-0.08	0.67	306	3/6
you	p	0.046	0.124	0.014	-0.08	0.37	289	3/6
you	t	0.072	0.182	0.015	-0.11	0.40	322	5/6
you	d	0.045	0.129	0.015	-0.08	0.35	264	4/6
you	q	0.043	0.146	0.029	-0.10	0.29	75	3/5
you	s	0.049	0.131	0.015	-0.08	0.37	271	4/5
you	r	0.053	0.180	0.016	-0.13	0.29	284	6/6
you	l	0.030	0.159	0.016	-0.13	0.19	266	6/6

Supplemental Table 4. Dependencies between signals involving the same concept. The effect is the genealogically balanced mean change in probability of finding the first symbol given that the second is present (as estimated by the mixed model). Only entries with q-values smaller than 0.05 shown. See Materials & Methods for further details.

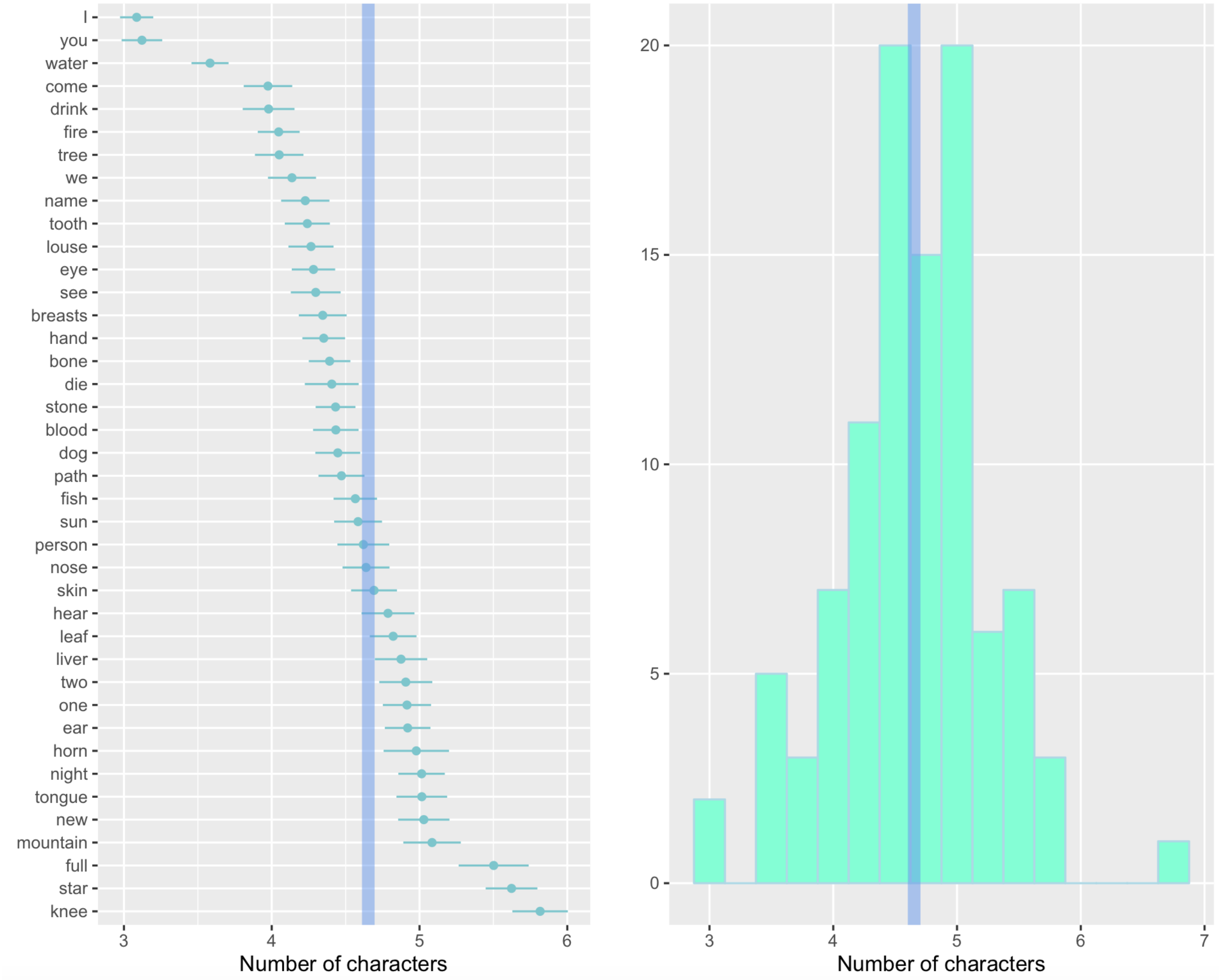
Concept	Symb.1	Symb.2	Effect	Fam. tested
bone	y	k	-0.04	298
bone	k	y	-0.14	298
breasts	h	m	-0.04	237
breasts	u	a	-0.16	314
breasts	u	m	-0.10	309
breasts	a	u	-0.16	314
breasts	a	m	0.18	317
breasts	a	r	0.11	285
breasts	m	h	-0.10	237
breasts	m	u	-0.08	309
breasts	m	a	0.12	317
breasts	r	a	0.03	285
dog	s	t	-0.08	281
dog	t	s	-0.04	281
full	b	p	-0.17	175
full	p	b	-0.21	175
I	b	t	0.02	264
I	s	u	-0.02	265
I	t	b	0.04	264
I	u	s	-0.07	265
knee	k	q	-0.19	71
knee	o	u	-0.28	273
knee	q	k	-0.22	71
knee	u	o	-0.29	273
leaf	l	b	0.10	217
leaf	b	l	0.09	217
leaf	b	p	-0.18	226
leaf	p	b	-0.21	226
name	o	i	-0.06	290
name	i	o	-0.12	290
nose	a	n	0.05	329
nose	a	u	-0.09	321
nose	n	a	0.05	329
nose	n	u	-0.05	319
nose	u	a	-0.09	321
nose	u	n	-0.06	319
one	n	t	-0.07	338
one	t	n	-0.06	338
tongue	E	e	-0.17	142
tongue	e	E	-0.16	142
tooth	b	m	0.03	272
tooth	m	b	0.02	272
we	l	n	-0.04	257
we	n	l	-0.19	257
we	n	p	-0.06	279
you	d	t	-0.04	253
you	r	o	0.02	254
you	u	o	-0.10	285
you	o	r	0.15	254
you	o	s	-0.04	252
you	o	u	-0.11	285
you	t	d	-0.04	253

Supplemental Table 5. Analysis of word-initial position bias. Bias measure how more or less frequently the symbol appears in word initial position for that concept. Lineages counts how many lineages had at least one language for which the analysis could be performed. See Materials & Methods for more details.

Concept	Symb.	Bias	Lineages
bite	k	0.20	42
bone	k	0.09	162
breasts	u	-0.06	185
breasts	m	0.05	152
ear	k	0.07	159
fish	a	0.05	249
full	b	0.11	81
full	p	0.12	100
horn	r	-0.23	82
horn	k	0.15	115
knee	o	0.10	177
knee	p	0.09	104
knee	k	0.07	177
knee	q	0.19	35
leaf	l	-0.14	120
one	n	-0.07	175
red	r	-0.24	28
tongue	l	-0.09	160

Supplemental Table 6. Estimated parameters (β), genealogical balanced mean probability difference (diff) and P-values for the distance to nearest neighbor (dnn) and population model, displayed only for the signals and variables that reached significance at $\alpha = 0.05$. See main text for details.

		Distance to nearest neighbor			Population		
Concept	Symb.	β	diff. (0-1000)	P-value	β	diff. (1-10000)	P-value
stone	t	-0.591	-0.296	0.009	-	-	-
full	p	-0.542	-0.444	0.005	-	-	-
dog	s	-0.441	-0.182	0.046	0.787	0.057	$< 10^{-3}$
tongue	E	-0.357	-0.343	0.034	-	-	-
knee	o	-0.263	-0.271	0.030	-	-	-
knee	u	-0.259	-0.249	0.023	-	-	-
nose	n	-0.244	-0.196	0.036	-	-	-
fish	a	-	-	-	1.009	0.176	$< 10^{-3}$
knee	p	-	-	-	-1.087	-0.121	$< 10^{-3}$
leaf	b	-	-	-	0.574	0.055	0.007
leaf	p	-	-	-	-0.506	-0.052	0.042
name	i	-	-	-	-0.420	-0.077	0.008
one	t	-	-	-	-0.575	-0.063	0.002
star	z	-	-	-	0.864	0.054	0.049
tongue	e	-	-	-	-0.358	-0.059	0.028
tongue	l	-	-	-	1.126	0.175	$< 10^{-3}$



Supplemental Figure 1. On the left, genealogically balanced average of the number of characters for each of the 40 concepts with most coverage in ASJP. The horizontal bars represent approximate 95% CI for the average. On the right, distribution of the genealogically balanced average for all of the concepts in ASJP. In both graphs, the vertical blue bar represents the mean value across all concepts in ASJP.