Ultraconserved words point to deep language ancestry across Eurasia


It is advisable to refer to the publisher’s version if you intend to cite from the work.
Published version at: http://dx.doi.org/10.1073/pnas.1218726110
To link to this article DOI: http://dx.doi.org/10.1073/pnas.1218726110

Publisher: National Academy of Sciences

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

CentAUR
Central Archive at the University of Reading

Reading's research outputs online
Ultraconserved words point to deep language ancestry across Eurasia

Mark Page1,b,d, Quentin D. Atkinson3, Andreea S. Calude6, and Andrew Meadea

1School of Biological Sciences, University of Reading, Reading, Berkshire RG6 6AS, United Kingdom; 2Santa Fe Institute, Santa Fe, NM 87501; 3School of Psychology, University of Auckland, Auckland 1142, New Zealand; and 4Linguistics Programme, University of Waikato, Hamilton 3240, New Zealand

Edited* by Colin Renfrew, University of Cambridge, Cambridge, United Kingdom, and approved April 15, 2013 (received for review October 31, 2012)

The search for ever deeper relationships among the World’s languages is bedeviled by the fact that most words evolve too rapidly to preserve evidence of their ancestry beyond 5,000 to 9,000 y. On the other hand, quantitative modeling indicates that some “ultra-conserved” words exist that might be used to find evidence for deep linguistic relationships beyond that time barrier. Here we use a statistical model, which takes into account the frequency with which words are used in common everyday speech, to predict the existence of a set of such highly conserved words among seven language families of Eurasia postulated to form a linguistic superfamily that evolved from a common ancestor around 15,000 y ago. We derive a dated phylogenetic tree of this proposed superfamily with a time-depth of ~14,450 y, implying that some frequently used words have been retained in related forms since the end of the last ice age. Words used more than once per 1,000 in everyday speech were 7- to 10-times more likely to show deep ancestry on this tree. Our results suggest a remarkable fidelity in the transmission of some words and give theoretical justification to the search for features of language that might be preserved across wide spans of time and geography.

cultural evolution | phylogeny | historical linguistics

The English word brother and the French frère are related to the Sanskrit bhūtṛ and the Latin frater, suggesting that words as mere sounds can remain associated with the same meaning for millennia. But how far back in time can traces of a word’s genealogical history persist, and can we predict which words are likely to show deep ancestry?

These questions are central to understanding language evolution and to efforts to identify linguistic superfamilies uniting the world’s languages (1–5). Evidence for proposed superfamilies—such as Amerind (6), linking most of the language families of the New World, and Nostratic (7–9) and Eurasiatic (3, 4, 10), linking the major language families of Eurasia—is often based on the identification of putative “cognate” words (analogous to homology in biology), the sound and meaning correspondences of which are thought to indicate that they derive from common ancestral words. Such evidence is often criticized for two reasons. First, most words are thought to suffer from too much semantic and phonetic erosion to allow secure identification of true cognates beyond 5,000 to 9,000 y (11, 12), and second, even if a number of apparent cognates can be identified, proponents of long-range relationships have been unable to provide statistical verification that the resemblances they have found are beyond what would be expected by chance between unrelated languages (11, 12). Where statistical tests have been used (9, 13), the results have been inconclusive because of the difficulty of establishing secure null models that estimate the number of resemblances expected to arise by chance.

Both objections can be overcome if it can be shown that: (i) a class of words exists whose members’ sound-meaning correspondences are expected to last long enough to retain traces of their ancestry between language families separated by thousands of years; and (ii) these ultraconserved words can be predicted a priori and independently of their sound correspondences to other words. Regarding the former, we have shown that most


The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

†To whom correspondence should be addressed. E-mail: m.page@reading.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1218726110/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1218726110

PNAS Early Edition | 1 of 6
as to the form of the word used by the common ancestor or proto-language of a given language family to denote a given meaning. These words are reconstructed by first identifying cognate words among the languages of a given family and then, because cognate words derive from a common ancestral word, working back in time to reconstruct the probable features of that shared ancestral form. Cognate relationships are recognized by patterns of shared sounds among pairs of words and by establishing regular patterns of sound change or “sound correspondences” among the contemporary, and sometimes “fossil” languages of a given language family. For example, the Latin *pater* is judged cognate to the English *father* on grounds of widely attested *p*→*f* and *t*→*th* transitions that occurred in the lineage leading to Germanic but not other Indo-European languages.

We recorded the proposed proto-words in the LWED for each of the 200 meanings in the Swadesh fundamental vocabulary list (19, 20), doing so separately for each of the seven language families in our sample. Often, linguists propose more than one proto-word for a given meaning, which can reflect synonyms in the proto-language or, more likely, uncertainty as to which of the words used among a language family’s extant languages are most likely to be cognate to the ancestral word. At the other extreme, for 12 meanings from the Swadesh list the LWED linguists could not reconstruct proto-words for more than two of the seven families, so these meanings were excluded from further analysis as not providing useful information for distinguishing relationships among the seven language families (deleting these meanings does not affect our results).

This process left 188 word-meanings for which one or more proto-words had been reconstructed for at least three language families (SI Text). We recorded all of these words, yielding 3,804 different reconstructed proto-words for the 188 × 7 = 1,316 possible pairings that arise for the 188 meanings among the seven language families. The modal number of reconstructed proto-words per meaning per language family is 1 (median = 2, mean = 2.89 ± 2.81, SD), and ranges from 1 to 26 (Fig. S1).

**Interfamily Cognates and Cognate Class Size.** For each of the proto-words, we searched among the proto-words for that meaning in the other language families to identify those that the LWED proposed as cognate between language families. Conventional comparative linguistic practice seeks to establish a set of “proven” cognates in garnering evidence for the existence of language families. However, at the time-depths interfamily cognates represent, the usual information—shared sounds and detection of regular sound correspondences—is often limited, making these cognates difficult to detect and susceptible to chance resemblances. We therefore adopt a statistical approach that does not depend upon individual cognates being proven. Instead, we treat each cognate proposal as a binary random variable subject to error, and seek evidence for regularities predicted to emerge when the set of proposed cognates derived from the 3,804 proto-words is taken as a whole.

We initially screened the proposed cognates, retaining only those in which the words kept the same reconstructed meaning across the language families, and we required a two-way correspondence in the judgement (SI Text). For example, the LWED proposes that the proto-Uralic form *to-nëe*, meaning “second,” and the proto-Kartvelian form *r̄ubu*, meaning “twins,” are both cognate to the proto-Indo-European form *dówu* and the proto-Altaic form *tjubu*, both of which mean “two.” Our “same meaning” criterion allows us to accept the proto-Indo-European and proto-Altaic proposals, but we exclude the proto-Uralic and proto-Kartvelian forms as cognates.

We then define the cognate class size for a given vocabulary item (meaning) as the number of language families whose proto-words for that item are hypothesized as cognate. Cognate class size can range from one, indicating a proto-word that is not cognate to the proto-words of any other language family, to seven, for a proto-word cognate across all seven language families. Larger cognate class sizes indicate words likely to be of greater antiquity, their forms having remained cognate across a larger number of language families. Where proto-words were not reconstructed for a language family we adopted the conservative view that the missing proto-words were not cognate to the other proto-words for that meaning in the different language families. The average cognate class size is 2.3 ± 1.1 (SD), with an observed range of 1–7 (mode = 2, median = 1.54) (Fig. 2).

**Predicting Cognate Class Size.** The positive skew to the distribution of cognate class sizes fits with our expectations, given the distribution of word half-lives (14): most lexical items have short linguistic half-lives of just a few thousand years, but a smaller set evolves slowly enough to remain cognate across the time-depths that separate language families (14).

If this reasoning is correct, we expect that words with larger cognate class sizes will be predictable from their rates of lexical replacement [the rate at which a word is replaced by a new noncognate word (14,15)], and from their frequency-of-use in
everyday speech. To test this prediction, we assembled for each of the 200 vocabulary items in the Swadesh list information on its rate of lexical replacement within the Indo-European language family (14), its generalized frequency-of-use in the worldwide sample reported previously (16) (Materials and Methods), and part of speech (Table S1).

We find that rates of lexical replacement predict the likelihood that a word will be judged as cognate among the seven Eurasian language families: words with slower rates of replacement have larger cognate class sizes, indicating older, more deeply retained words (Fig. 3A) \( r = -0.43, P < 0.001 \). This relationship holds separately for each of the 21 possible pairs of language families: in each pair the proto-words judged cognate between the two families have slower average rates of lexical replacement than the proto-words judged noncognate (sign test, \( P < 0.001 \)).

Generalized frequency-of-use, along with part of speech, is also a significant predictor of cognate class size (Fig. 3B) \( r = 0.48, P < 0.001 \). For a range of words used at low frequencies, maximum cognate class size remains stable at around two (most have the minimum cognate class size of one), but as frequency-of-use increases above a threshold, the size of the cognate class steadily increases. This result suggests that, consistent with their short estimated half-lives, infrequently used words typically do not exist long enough to be deeply ancestral, but that above the threshold frequency words gain greater stability, which then translates into larger cognate class sizes. Generalized frequency-of-use does not contribute to the prediction of cognate class size after controlling for rates of lexical replacement (\( P = 0.253 \)), consistent with the view that frequency-of-use acts on cognate class size via its influence on the rate of lexical replacement.

In Fig. 3, rapidly evolving words (lower frequency or higher lexical replacement rate) act as controls for the slowly evolving words to estimate the likelihood of chance sound correspondences. In both cases, the rapidly evolving words tend to have cognate class sizes less than two (i.e., not cognate to any other proto-word), showing that the influence of chance resemblances on the cognacy judgements is low.

A rule-of-thumb emerges (Fig. 3B) that words used more than around once per 1,000 in everyday speech evolve slowly enough to have a high chance of being judged cognate among more than two of the language families; this might equate to around 16 uses per day per speaker of these high-frequency words (21). Twenty-three meanings had cognate class sizes of four or more (Table 1). Our expectation is that these highly conserved words will be
those with unusually high frequencies of use, particularly among the numerals, pronouns, and special adverbs (14). Words used more than once per 1,000 spoken words are overrepresented on this list (χ² = 24.29, P < 0.001), as are pronouns and adverbs (χ² = 26.1, P < 0.0001 and χ² = 14.5, P = 0.003, respectively). The odds ratio comparing the probability that a word has a cognate class size of four or more, given that it is frequently used (<0.01), to the probability obtained ignoring frequency is 10 (P < 0.001; controlling for part of speech it is 7.5, P < 0.001): frequently used words are at least seven-times more likely to be judged cognate.

By controlling for the likelihood of chance sound associations, these analyses give us confidence that words such as “thou,” “I,” “who,” “not,” “that,” “to give,” and “we” are probably ancient, being cognate among four or more language families. A few words, including “bark,” are infrequently used today but nevertheless appear conserved. The numeral words, despite having some evidence in-
enough to identify cognate relationships among the language families of Eurasia.

Our ability to predict these words independently of their sound correspondences dilutes the usual criticisms leveled at such long-range linguistic reconstructions, that proto-words are unreliable or inaccurate, or that apparent phonetic similarities among them reflect chance sound resemblances. Error in proposed proto-words would weaken the signals we detect and chance sound resemblances would arise just as often in infrequently as well as frequently used words; however, we find significantly more cognates, as predicted, among the frequently used and slowly evolving words.

Still, three kinds of criticisms might be directed at the long-range comparisons on which we base our analyses: (i) that proposed cognates might arise from borrowings; (ii) that historical linguists are more likely to declare frequently used proto-words cognate simply by virtue of their implied stability, and do so independently of their sound correspondences; and, (iii) that some categories of words are more likely by chance to appear cognate than others. We consider each of these possibilities (see also SI Text).

For borrowings systematically to affect our results, lexical items would have to have been exchanged so frequently among the many extant languages of two or more language families as to cause them to be reconstructed as the proto-words in both families. Alternatively, perhaps some of our cognate proto-words arise from words that were borrowed so early in the histories of two language families, and then retained in the descendant languages, as to become widespread among the contemporary languages of both. This process would only affect our results if such early adaptations were widespread and biased toward frequently used words, the stability of which made it likely that they would be retained in the many descendant languages. Instead, frequently used words are less likely to be adopted: recent data (28) show that rates of borrowing among the words in the Swadesh list are generally low, and especially so for the 23 words of Table 1.

For these reasons we also think it unlikely that the correspondence between our proposed tree and geography merely reflects the effects of areal diffusion or borrowing. A structure of the topology we derive in Fig. 4.4 supports these arguments by placing language families that are geographical neighbors in distinct regions of the tree. For example, the Altaic language family includes modern-day Turkish, which is surrounded by Indo-European languages, and yet proto-Altaic is placed distantly to proto-Indo-European. Similarly, proto-Dravidian and especially proto-Kartvelian are distant to proto-Indo-European and proto-Altaic, despite their likely central Asian origins.

Perhaps the LWED linguists are more likely to find links between high-frequency words or words that evolved more slowly within families, such as Indo-European, simply by virtue of their implied stability. We cannot rule out this bias, but note there are some relatively high-frequency/stable words (e.g., “to say,” “day,” and “to know,” along with the number words) with cognate class sizes of two or less, and some infrequently used words are judged to be conserved (e.g., “bark,” “ashes,” and “worm”). This finding shows that if a bias exists, it does not mechanically overrule other signals in the data pointing either toward cognacy or the lack of it. In addition, the LWED proposes many more possible proto-words for the less-frequently used meanings (reflecting the greater variety of words for these meanings within and among languages). It then examines all possible pairs of proto-words between two language families for evidence of sound correspondences that might imply a cognate link. The large number of possible comparisons means that just by chance, one expects more cognate links to be found among the infrequently used meanings: but we find the opposite.

Are some categories of words more likely to appear cognate by chance? None of the words in Table 1 are closed-class words of simple phonology (“thou,” “I,” “not,” “that,” “we,” “who,” “this,” “what,” “ye”) whose short length might mean that chance resemblances between their proto-words are more likely. Comparative linguists are aware of this potential source of bias and often avoid reconstructing proto-words for these closed-class words. Indeed, all 12 meanings that we excluded from our analyses because the LWED linguists could not derive proto-words for them are
closed-class words of this type. Removing the nine closed-class words from Table 1 does not change any of our conclusions.

Our results support the findings (14) that human language can achieve a remarkable degree of replication fidelity among its highly used words, and especially so for some parts of speech. If the Eurasiatic superfamily is around 15-ky old, then traces of the sounds from a predictable subset of words have remained associated with their particular meanings independently in separate branches of this superfamily since the end of the last ice age. This finding is all of the more surprising given that words are culturally transmitted replicators (27), passed many thousands of times from speaker to speaker every generation, and subject to the potentially corrupting influences of competing words, borrowings, and sound production errors.

Proposals that link large numbers of the world’s languages into linguistic superfamilies are frequently criticized (11–13, 29), but this view needs revising (see, for example, refs. 30–32). Our statistical model overcomes objections to the identification and existence of deep cognate relationships by providing a quantitative framework for expecting such deep links in a subset of vocabulary items, and lends a theoretical plausibility to the search for further candidate words unifying other linguistic families.

Materials and Methods

Languages of the World Etymological Database. The LWED is part of the Tower of Babel project, a collaboration founded by the late Sergei Starostin (Si Text) (http://starling.rinet.ru/cgi-bin/main.cgi) and affiliated with the Evolution of Human Languages project at the Santa Fe Institute (http://ehl.santafe.edu/main.html).

Generalized Frequency-of-Use. A vocabulary item’s generalized frequency-of-use is calculated as the logarithm of its mean frequency in 17 languages from six language families, plus Basque and the Creole Tok Pisin (16). This measure correlates 0.99 with the first principal component of these same frequencies (16).

Phylogenetic Inference. We estimated a posterior distribution of phylogenetic trees from a MCMC procedure (Si Text) applied to the pairs of distances between languages on phylogenetic trees. The Markov chain proposes a new tree and branch-lengths each iteration of the chain, and then evaluates the likelihood of the distances that tree implies. We estimate the likelihood of a distance between a pair of languages i and j by evaluating

\[ L_i = \prod_{k=1}^{m} \sum_{r_{i,j} = 0}^{\infty} \sum_{k=1}^{n} \frac{r_{i,j}^k \gamma_{i,j}^k P_{i,j}}{k!} \]

for a given t or unknown time, where \( P_{i,j} = (1 - e^{-\alpha_{ij}}) \) and \( \gamma_{i,j} = (e^{-\alpha_{ij}}) \), m corresponds to words in the Swadesh list that we scored as not cognate between the two language families, \( n = (m + 1) \) counts the words scored as cognate, \( \alpha_{ij} \) is the rate of change for the \( k^\text{th} \) word in units of lexical replacement per unit time, as estimated in the Indo-European languages (rates taken from ref. 14), and \( \gamma_{i,j} \) is the usual \( \gamma \)-rate heterogeneity (33) summed over four rate categories.

ACKNOWLEDGMENTS. This work was supported by grants from the Leverhulme Trust and from the European Research Council (to M.P.)