Ultraconserved words point to deep language ancestry across Eurasia

Mark Pagel[a,b,1], Quentin D. Atkinson[c], Andreea S. Calude[d], and Andrew Meade[a]

[a]School of Biological Sciences, University of Reading, Reading, Berkshire RG6 6AS, United Kingdom; [b]Santa Fe Institute, Santa Fe, NM 87501; [c]School of Psychology, University of Auckland, Auckland 1142, New Zealand; and [d]Linguistics 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 "ultraconserved" 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 *bhrātr* and the Latin *frāter*, 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 words have about a 50% chance of being replaced by a new noncognate word [a word's linguistic half-life (14, 15)], roughly every 2,000–4,000 y, consistent with the belief that words lose traces of their ancestry quickly. However, some words, such as the numerals, pronouns, and special adverbs (e.g., I, you, here, how, not, there, what, two, five) are replaced far more slowly, with half-lives of once every 10,000, 20,000 or even more years (14, 15).

Usefully, these words can be predicted from information independent of their sounds. We showed in a sample of Indo-European languages that the frequency with which a word is used in everyday speech, along with its part of speech, can predict how rapidly words evolve, with frequently used words on average retained for longer periods of time (14). We have recently extended this result to include speakers from the Uralic, Sino-Tibetan, Niger-Congo, Altaic, and Austronesian families, in addition to Indo-European, plus the isolate Basque and the Creole Tok Pisin (16). Even in languages as widely divergent as these, we found that a measure of the average frequency of use predicted rates of lexical replacement as estimated in the Indo-European languages.

Taken together, these findings suggest that the way we use a core set of vocabulary words in everyday speech is a stable and shared feature of human discourse, and raises the possibility that words will evolve in other language families at rates similar to those found in the Indo-European languages, with frequency of word-use acting as the common causal factor. This provides a statistical framework for predicting—without recourse to sound correspondences—words likely to show deep ancestry among languages and even among language families whose relationships might extend well beyond 10,000 y.

We use this framework to predict words likely to be shared among the Altaic, Chukchi-Kamchatkan (sometimes called Chukotko-Kamchatkan or Chukchee-Kamchatkan), Dravidian, Eskimo (hereafter referred to as Inuit-Yupik) (*SI Text*), Indo-European, Kartvelian, and Uralic language families. These seven language families are hypothesized to form an ancient Eurasiatic superfamily that may have arisen from a common ancestor over 15 kya (17), and whose languages are now spoken over all of Eurasia (Fig. 1 and *SI Text*).

Results

Proto-Words. The Languages of the World Etymological Database, part of the Tower of Babel project (LWED) (18) (*Materials and Methods* and *SI Text*), records reconstructed proto-words for language families from around the world. Proto-words are hypotheses

Author contributions: M.P., Q.D.A., A.S.C., and A.M. performed research; M.P. and A.M. contributed new reagents/analytic tools; M.P., Q.D.A., A.S.C., and A.M. analyzed data; and M.P., Q.D.A., and A.S.C. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: m.pagel@reading.ac.uk.

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

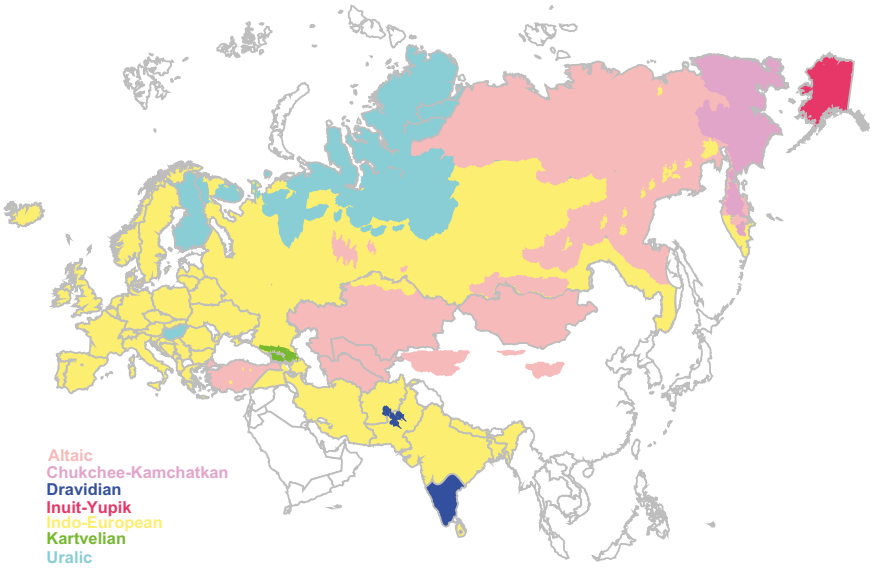

Fig. 1. Map showing approximate regions where languages from the seven Eurasiatic language families are spoken. The color-shaded areas should be treated as suggestive only, as current language ranges will not necessarily correspond to original homelands, and language boundaries will often overlap. For example, the Indo-European language Swedish is spoken along with the Uralic Finnish in southern Finland (map source: refs. 10, 16 and 34) (*SI Text*).

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 $188 \times 7 = 1{,}316$ 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-ńće, meaning "second," and the proto-Kartvelian form *ṭqub-, meaning "twins," are both cognate to the proto–Indo-European form *duwo and the proto-Altaic form *tu̯bu, 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

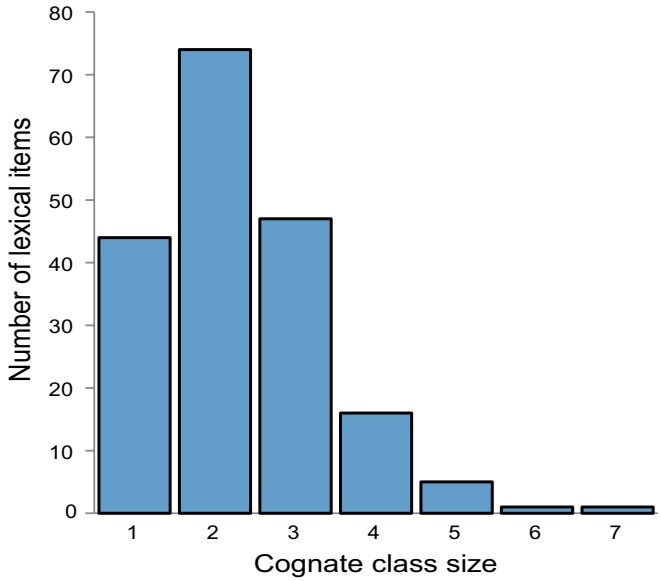

Fig. 2. Cognate class sizes. The number of meanings (n = 188 meanings, i.e., excluding the 12 meanings discussed in the text) (*SI Text*) with cognate class sizes ranging from one (not cognate to any other language family) to seven (cognate across all seven language families), mean = 2.3 ± 1.1 (SD).

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 Eurasiatic language families: words with slower rates of replacement have larger cognate class sizes, indicating older, more deeply retained words (Fig. 3*A*) ($r = -0.43$, $P < 0.001$). 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. 3*B*) ($r = 0.48$, $P < 0.001$). 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. 3*B*) 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

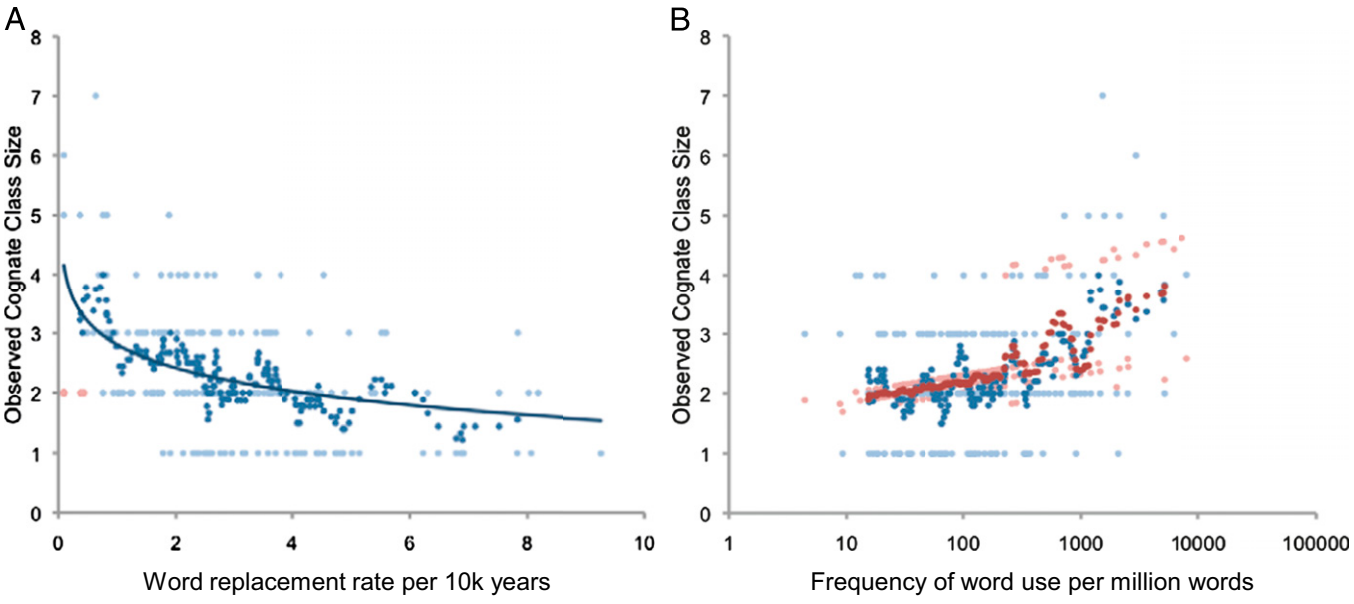

Fig. 3. Rates of lexical replacement (*A*) and word-use frequencies (*B*) predict cognate class sizes among the seven Eurasiatic language families. (*A*) Cognate class size predicted from the rate of lexical replacement as measured in the Indo-European languages [n = 188 lexical items, $r = -0.43$, $P < 0.001$; excluding number terms (see text) the correlation increases to $r = -0.55$]. Rates record the expected number of replacements by a new unrelated word per 1,000 years (14); cognate class size is the number out of the seven families for which the proto-word of a given meaning is ancestrally shared; the correlation is fitted to the raw data; smoothed data (darker symbols) are based on a running mean with a window width of 10. (*B*) Cognate class size predicted from a regression model combining frequency of word use and part of speech (see text). The regression is calculated on raw data (blue), smoothed data as in *A* ($r = 0.48$, $P < 0.001$). The trend in *B* is unchanged if we use the principal component factor scores (*Materials and Methods*) in place of mean frequencies.

Table 1. Twenty-three words with cognate class sizes of four or more among the Eurasiatic language families

Meaning	Cognate class size*	I-E rate†	Half-life 1,000s of years	Frequency of use‡	Part of speech
Thou	7	0.064	10.83	2,524	Pronoun
I	6	0.009	77	4,332	Pronoun
Not	5	0.082	8.45	7,602	Adverb
That	5	0.188	3.69	5,846	Adjective
We	5	0.037	18.73	2,956	Pronoun
To give	5	0.076	9.12	1,606	Verb
Who	5	0.009	77	1,172	Pronoun
This	4	0.218	3.18	11,185	Adjective
What	4	0.069	10.04	3,058	Adverb
Man/male	4	0.338	2.05	2,800	Noun
Ye	4	0.132	5.25	1,459	Pronoun
Old	4	0.253	2.74	746	Adjective
Mother	4	0.236	2.94	717	Noun
To hear	4	0.235	2.95	680	Verb
Hand	4	0.082	8.45	658	Noun
Fire	4	0.175	3.96	398	Noun
To pull	4	0.453	1.71	279	Verb
Black	4	0.191	3.62	135	Adjective
To flow	4	0.34	2.04	91	Verb
Bark	4	0.379	1.82	49	Noun
Ashes	4	0.265	2.62	23	Noun
To spit	4	0.204	3.38	23	Verb
Worm	4	0.216	3.19	21	Noun

*Defined as the number (of seven) of Eurasiatic language families that are reconstructed as cognate for the word used to convey the meaning shown.
†The rate of lexical replacement measured in number of expected new or unrelated words per 1,000 y and rates of replacement expressed as "half-lives" or the expected time until a word has a 50% chance of being replaced by a new noncognate word (14).
‡The frequency of use per million based on mean of 17 languages from six language families and the two isolates (16).

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 ($\chi^2 = 24.29$, $P < 0.001$), as are pronouns and adverbs ($\chi^2 = 26.1$, $P < 0.0001$ and $\chi^2 = 14.5$, $P = 0.003$, 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 ($f >$ 1,000), 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 of the slowest rates of lexical replacement in the Indo-European languages, have cognate class sizes of only two and do not appear in Table 1. Our conservative coding might have contributed to this, but number words are known to change among language families. These words can be invented independently (22), or because of their importance to communication and administration, they might be replaced *en bloc* and possibly at times of political or social unrest, as has been true historically of words for months of the year.

Phylogenetic Tree of the Eurasiatic Language Superfamily. We can use the cognate proposals along with rates of lexical replacement to estimate a dated phylogenetic tree of these Eurasiatic language families. We first recorded for each word in the Swadesh list whether the proto-words of a pair of language families were scored as cognate (1) or not (0). We produced these lists for all 21 pairs of language families, and then associated with each of the cognate proposals in the list the independently derived rate of lexical replacement for the corresponding meaning (14).

We used these data to infer a posterior distribution of phylogenetic trees of the seven families from a Markov chain Monte Carlo (MCMC) approach that simultaneously incorporated uncertainty in the data and uncertainty in the timings of internal dates on the tree (*SI Text* and Table S2). The time-distance between any pair of languages on the tree is constrained by having to satisfy the pairwise distributions of cognate proposals, given the rates of lexical replacement.

The same consensus unrooted tree emerged from five independent MCMC runs (Fig. 4*A*) and displays three sets of relationships: a Central and Southern Asian grouping of the Kartvelian and Dravidian language families, a northern and western European grouping of the Indo-European and Uralic families, and an eastern grouping including the Altaic, Inuit-Yupik, and Chukchi-Kamchatkan families. The consensus topology is also the most frequently occurring topology in our posterior sample of trees, and 9 of the 10 most frequently occurring topologies place proto-Dravidian and proto-Kartvelian in this position outside of the others (Bayes Factor = 29, indicating strong support) (23). This finding agrees with suggestions (17) that the Kartvelian and Dravidian language families are related to, but fall outside, a core group of Eurasiatic languages comprising proto-Indo-European and proto-Uralic in the west and proto-Altaic, proto–Inuit-Yupik, and proto–Chukchi-Kamchatkan in the east.

Genetic data suggest that Dravidian populations could represent an early expansion from Central to Southern Asia that almost certainly occurred before the arrival of the Indo-Europeans (24). Rooting the tree at the midpoint along the branch leading to proto-Dravidian (Fig. 4*B*) yields an age for the origin of the Eurasiatic superfamily of 14.45 ± 1.75 kya [95% confidence interval (CI) = 11.72–18.38 kya]. Consistent with the Dravidian expansion being ancient, the tree makes proto-Dravidian older than proto–Indo-European [although some scholars think that the common ancestor of contemporary Dravidian languages is younger (25)]. An alternative root, placed along the branch to proto-Kartvelian, produces a slightly older tree (15.61 ± 2.29 kya, 95% CI = 11.72–20.40 kya; agreement between two lower 95% CIs is coincidence).

Posterior support at internal nodes of the tree is low, as we might expect of a linguistic tree of this age, but all exceed chance expectations (*SI Text*) and the internal topology does not affect our estimates of the age of the superfamily. All inferred ages must be treated with caution but our estimates are consistent with proposals linking the near concomitant spread of the language families that comprise this group to the retreat of glaciers in Eurasia at the end of the last ice age ~15 kya (4, 17). The 95% CIs around the root-age are consistent with the initial separation of these families occurring before the development of agriculture beginning ~11 kya (26).

Discussion

The key question of long-range or deep reconstruction in historical linguistics is whether one can expect enough phonetic trace to survive time-depths exceeding what has become the informal limit of 8,000 to 9,000 y, over which it is considered that cognacy relationships can be judged reliably. Our previous work (14, 15, 27) providing quantitative estimates of the rates at which words are replaced by new noncognate words (as judged by the lack of sound correspondence) indicated that the answer to that question is yes for a small but nevertheless predictable subset of words. Here we have applied that methodology to identify a priori the words we can expect to be retained for periods long

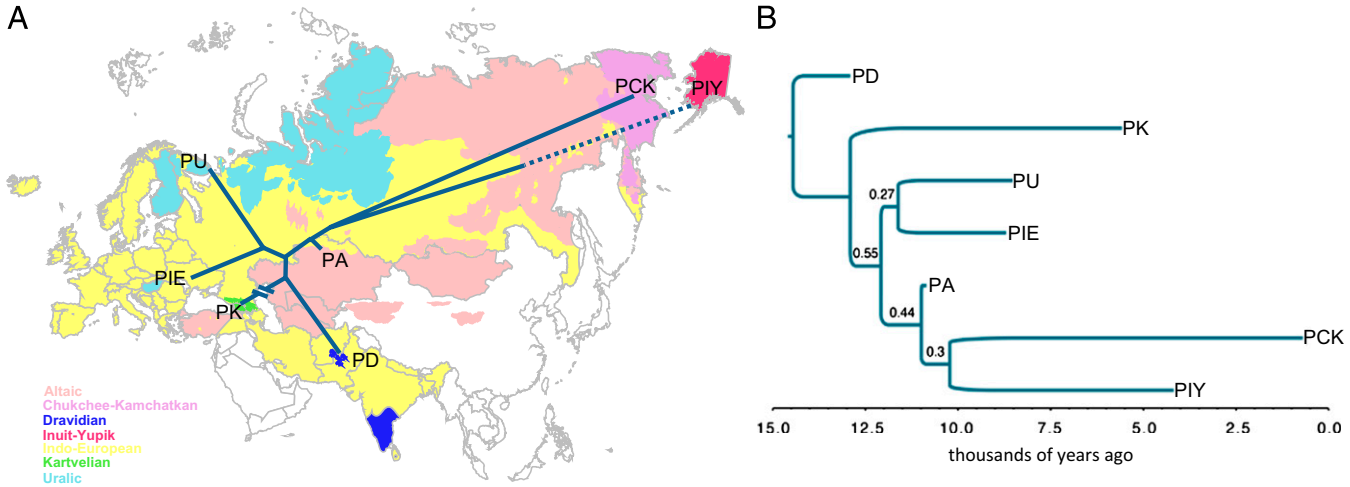

Fig. 4. Consensus phylogenetic tree of Eurasiatic superfamily (*A*) superimposed on Eurasia and (*B*) rooted tree with estimated dates of origin of families and of superfamily. (*A*) Unrooted consensus tree with branch lengths (solid lines) shown to scale and illustrating the correspondence between the tree and the contemporary north-south and east-west geographical positions of these language families. Abbreviations: P (proto) followed by initials of language family: PD, proto-Dravidian; PK, proto-Kartvelian; PU, proto-Uralic; PIE, proto–Indo-European; PA, proto-Altaic; PCK, proto–Chukchi-Kamchatkan; PIY, proto–Inuit-Yupik. The dotted line to PIY extends the inferred branch length into the area in which Inuit-Yupik languages are currently spoken: it is not a measure of divergence. The cross-hatched line to PK indicates that branch has been shortened (compare with *B*). The branch to proto-Dravidian ends in an area that Dravidian populations are thought to have occupied before the arrival of Indo-Europeans (see main text). (*B*) Consensus tree rooted using proto-Dravidian as the outgroup. The age at the root is 14.45 ± 1.75 kya (95% CI = 11.72–18.38 kya) or a slightly older 15.61 ± 2.29 kya (95% CI = 11.72–20.40 kya) if the tree is rooted with proto-Kartvelian. The age assumes midpoint rooting along the branch leading to proto-Dravidian (rooting closer to PD would produce an older root, and vice versa), and takes into account uncertainty around proto–Indo-European date of 8,700 ± 544 (SD) y following ref. 35 and the PCK date of 692 ± 67 (SD) y ago (*SI Text*).

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 adoptions 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. The structure of the topology we derive in Fig. 4*A* 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? Nine 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 uniting 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

$$L_{ij} = \prod_{k=1}^{m} \sum_{i=1}^{4} \gamma_i P_{k0} \times \prod_{k=m+1}^{n} \sum_{i=1}^{4} \gamma_i P_{k1} \quad [1]$$

for a given t or unknown time, where $P_{k0} = (1 - e^{-r_k t})$ and $P_{k1} = (e^{-r_k t})$, 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, r_k is the rate of change for the k^{th} word in units of lexical replacement per unit time, as estimated in the Indo-European languages (rates taken from ref. 14), and γ_i is the usual γ-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.).

1. Bomhard A (2008) *Reconstructing Proto-Nostratic: Comparative Phonology, Morphology, and Vocabulary* (Brill, Boston, MA).
2. Greenberg J (1957) *Essays in Linguistics* (Univ of Chicago Press, Chicago).
3. Greenberg J (2000) *The Eurasiatic Language Family: Indo-European and Its Closest Relations. Volume I: Grammar* (Stanford Univ Press, Stanford, CA).
4. Greenberg J (2002) *Indo-European and its Closest Relatives: The Eurasiatic Language Family Volume II: Lexicon* (Stanford Univ Press, Stanford, CA).
5. Ruhlen M (1994) *The Origin of Language: Tracing the Evolution of the Mother Tongue* (Wiley, New York).
6. Greenberg J (1987) *Language in the Americas* (Stanford Univ Press, Stanford, CA).
7. Illich-Svitych V (1989) in *Reconstructing Languages and Cultures*, ed Shevoroshkin V (Universitatsverlag Brockmeyer, Bochum), pp 125–176.
8. Dolgopolsky A (1986) in *Typology, Relationship and Time*, eds Shevoroshkin V, Markey T (Karoma, Ann Arbor, MI), pp 27–50.
9. Oswalt R (1998) in *Nostratic: Sifting the Evidence*, eds Joseph S, Joseph B (John Benjamins, Amsterdam, The Netherlands), pp 199–216.
10. Ruhlen M (1987) *A Guide to the World's Languages, Volume 1: Classification* (Stanford Univ Press, Stanford, CA).
11. Ringe D (1995) 'Nostratic' and the factor of chance. *Diachronica* 12:55–74.
12. Campbell L (2008) in *Language Classification: History and Method* (Cambridge Univ Press, Cambridge).
13. Ringe D (1998) in *Nostratic: Sifting the Evidence*, eds Joseph S, Joseph B (John Benjamins, Amsterdam, The Netherlands), pp 143–198.
14. Pagel M, Atkinson QD, Meade A (2007) Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* 449(7163):717–720.
15. Pagel M, Meade A (2006) in *Phylogenetic Methods and the Prehistory of Languages*, eds Clackson J, Forster P, Renfrew C (MacDonald Institute for Archaeological Research, Cambridge), pp 173–182.
16. Calude AS, Pagel M (2011) How do we use language? Shared patterns in the frequency of word-use across seventeen World languages. *Phil Trans Roy Soc B* 366:1567):1101–1107.
17. Bomhard A, Kerns J (1994) *The Nostratic Macrofamily* (Mouton de Gruyter, Amsterdam, The Netherlands).
18. Starostin SA, Bronnikov Y (1998–2009) 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. Also available at http://ehl.santafe.edu/main.html. Accessed March 12, 2012.
19. Swadesh M (1952) Lexicostatistic dating of prehistoric ethnic contacts. *Proc Am Philos Soc* 96(4):452–463.
20. Swadesh M (1955) Towards greater accuracy in lexicostatistic dating. *Int J Am Linguist* 21(2):121–137.
21. Mehl MR, Vazire S, Ramírez-Esparza N, Slatcher RB, Pennebaker JW (2007) Are women really more talkative than men? *Science* 317(5834):82.
22. Hurford J (1987) *Language and Number: The Emergence of a Cognitive System* (Blackwell, Oxford).
23. Raftery AE (1996) in *Markov Chain Monte Carlo in Practice*, eds Gilks WR, Richardson S, Spiegelhalter DJ (Chapman & Hall, London), pp 163–188.
24. Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian population history. *Nature* 461(7263):489–494.
25. Andronov MS (1964) Lexicostatistic analysis of the chronology of disintegration of Proto-Dravidian. *Indo Iran J* 7(2–3):170–186.
26. Renfrew C (1991) Before Babel, speculations on the origins of linguistic diversity. *Camb Archaeol J* 1(1):3–23.
27. Pagel M (2009) Human language as a culturally transmitted replicator. *Nat Rev Genet* 10(6):405–415.
28. Haspelmath M, Tadmor U (2009) *Loanwords in the World's Languages: A Comparative Handbook* (Mouton de Gruyter, Amsterdam, The Netherlands).
29. McMahon A, McMahon R (2005) *Language Classification by Numbers* (Oxford Univ Press, Oxford).
30. Vajda E (2010) in *The Dene-Yeniseian Connection*, eds Kari J, Potter BA (Univ of Alaska Press, Fairbanks, AK), pp 33–100.
31. Ruhlen M (1994) *On the Origin of Language: Studies in Linguistic Taxonomy* (Stanford Univ Press, Stanford, CA).
32. Dediu D, Levinson SC (2012) Abstract profiles of structural stability point to universal tendencies, family-specific factors, and ancient connections between languages. *PLoS ONE* 7(9):e45198.
33. Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J Mol Evol* 39(3):306–314.
34. Crystal D (2010) *Cambridge Encyclopedia of Language* (Cambridge Univ Press, Cambridge).
35. Gray RD, Atkinson QD (2003) Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426(6965):435–439.