Vol 449 | 11 October 2007 | doi:10.1038/nature06176

nature

Frequency of word-use predicts rates of lexical evolution throughout Indo-European history

Mark Pagel[1,2], Quentin D. Atkinson[1] & Andrew Meade[1]

Greek speakers say "*ουρά*", Germans "*schwanz*" and the French "*queue*" to describe what English speakers call a 'tail', but all of these languages use a related form of 'two' to describe the number after one. Among more than 100 Indo-European languages and dialects, the words for some meanings (such as 'tail') evolve rapidly, being expressed across languages by dozens of unrelated words, while others evolve much more slowly—such as the number 'two', for which all Indo-European language speakers use the same related word-form[1]. No general linguistic mechanism has been advanced to explain this striking variation in rates of lexical replacement among meanings. Here we use four large and divergent language corpora (English[2], Spanish[3], Russian[4] and Greek[5]) and a comparative database of 200 fundamental vocabulary meanings in 87 Indo-European languages[6] to show that the frequency with which these words are used in modern language predicts their rate of replacement over thousands of years of Indo-European language evolution. Across all 200 meanings, frequently used words evolve at slower rates and infrequently used words evolve more rapidly. This relationship holds separately and identically across parts of speech for each of the four language corpora, and accounts for approximately 50% of the variation in historical rates of lexical replacement. We propose that the frequency with which specific words are used in everyday language exerts a general and law-like influence on their rates of evolution. Our findings are consistent with social models of word change that emphasize the role of selection, and suggest that owing to the ways that humans use language, some words will evolve slowly and others rapidly across all languages.

Languages, like species, evolve by way of a process of descent with modification (Supplementary Table 1). The remarkable diversity of languages—there are about 7,000 known living languages[7]—is a product of this process acting over thousands of years. Ancestral languages split to form daughter languages that slowly diverge as shared lexical, phonological and grammatical features are replaced by novel forms. In the study of lexical change, the basic unit of analysis is the cognate. Cognates are words of similar meaning with systematic sound correspondences indicating they are related by common ancestry. For example, cognates meaning 'water' exist in English (water), German (*wasser*), Swedish (*vatten*) and Gothic (*wato*), reflecting descent from proto-Germanic (**water*).

Early lexicostatistical[8] studies of Malayo-Polynesian and Indo-European language families revealed that the rate at which new cognates arise varies across meaning categories[1,9]. More recently we have obtained direct estimates of rates of cognate replacement on linguistic phylogenies (family trees) of Indo-European and Bantu languages, using a statistical model of word evolution in a bayesian Markov chain Monte Carlo (MCMC) framework[10]. We found that rates of cognate replacement varied among meanings, and that rates for different meanings in Indo-European were correlated with their paired meanings in the Bantu languages. This indicates that variation in the rates of lexical replacement among meanings is not merely an historical accident, but rather is linked to some general process of language evolution.

Social and demographic factors proposed to affect rates of language change within populations of speakers include social status[11], the strength of social ties[12], the size of the population[13] and levels of outside contact[14]. These forces may influence rates of evolution on a local and temporally specific scale, but they do not make general predictions across language families about differences in the rate of lexical replacement among meanings. Drawing on concepts from theories of molecular[15] and cultural evolution[16–18], we suggest that the frequency with which different meanings are used in everyday language may affect the rate at which new words arise and become adopted in populations of speakers. If frequency of meaning-use is a shared and stable feature of human languages, then this could provide a general mechanism to explain the large differences across meanings in observed rates of lexical replacement. Here we test this idea by examining the relationship between the rates at which Indo-European language speakers adopt new words for a given meaning and the frequency with which those meanings are used in everyday language.

We estimated the rates of lexical evolution for 200 fundamental vocabulary meanings[8] in 87 Indo-European languages[6]. Rates were estimated using a statistical likelihood model of word evolution[10] applied to phylogenetic trees of the 87 languages (Supplementary Fig. 1). The number of cognates observed per meaning varied from one to forty-six. For each of the 200 meanings, we calculated the mean of the posterior distribution of rates as derived from a bayesian MCMC model that simultaneously accounts for uncertainty in the parameters of the model of cognate replacement and in the phylogenetic tree of the languages (Methods). Rate estimates were scaled to represent the expected number of cognate replacements per 10,000 years, assuming a 8,700-year age for the Indo-European language family[6]. Opinions on the age of Indo-European vary between approximately 6,000 and 10,000 years before present[19,20]. Using a different calibration would change the absolute values of the rates but not their relative values.

Figure 1a shows the inferred distribution of rate estimates, where we observe a roughly 100-fold variation in rates of lexical evolution among the meanings. At the slow end of the distribution, the rates predict zero to one cognate replacements per 10,000 years for words such as 'two', 'who', 'tongue', 'night', 'one' and 'to die'. By comparison, for the faster evolving words such as 'dirty', 'to turn', 'to stab' and 'guts', we predict up to nine cognate replacements in the same time period. In the historical context of the Indo-European language family, this range yields an expectation of between 0–1 and 43 lexical replacements throughout the ~130,000 language-years of evolution the linguistic tree represents, very close to the observed range in the

[1] School of Biological Sciences, University of Reading, Whiteknights, Reading, Berkshire, RG6 6AS, UK. [2] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA.

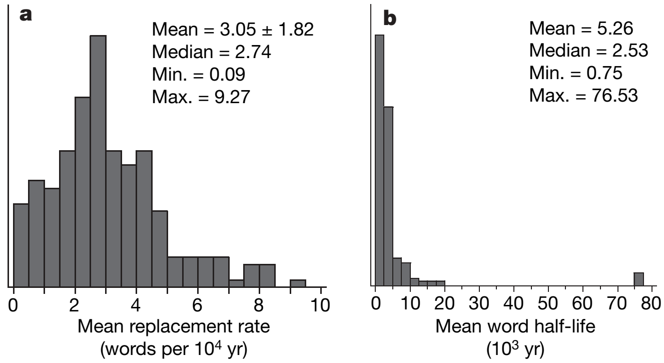

Figure 1 | Frequency plots for rates of lexical evolution in Indo-European across 200 fundamental vocabulary meanings. a, The mean estimated rate of cognate replacement for each meaning. **b,** The same rate distribution converted to word half-lives[10], or the time in which there is a 50% chance the word will be replaced by a different non-cognate form. The longest half-lives (76,530 years) are for meanings that show no change across Indo-European (Supplementary Information).

fundamental vocabulary of 1–46 distinct cognate classes among the different meanings. These rates can be converted to estimates of the linguistic half-life[10] (Methods), or the time in which there is a 50% chance the word will be replaced by a different non-cognate form. These times vary from 750 years for the fastest evolving words to over 10,000 years for the slowest (Fig. 1b).

We used spoken and written language corpus data from English[2], Spanish[3], Russian[4] and Greek[5] to measure the frequency of meaning-use (Supplementary Table 2). These languages sample from across the Indo-European language family (Supplementary Fig. 1), and their corpora were selected to provide large samples of language use (20–100 million words each). Figure 2 shows that the distribution of word-use frequencies in each language is highly skewed, such that most words are used relatively infrequently (fewer than 100 times per million words), with a small number of frequently used words (as often as 35,000 times per million words) accounting for most speech. Word-use frequencies are highly correlated among the four languages ($0.78 < r < 0.89$, mean $r = 0.84$; Supplementary Fig. 2), showing that words used at a high frequency in one language tend to be used at a high frequency in the other languages. Because the four languages span the Indo-European tree, this suggests that frequency values are representative of Indo-European language use, and that frequencies of meaning-use have been remarkably stable throughout Indo-European history (Supplementary Fig. 1).

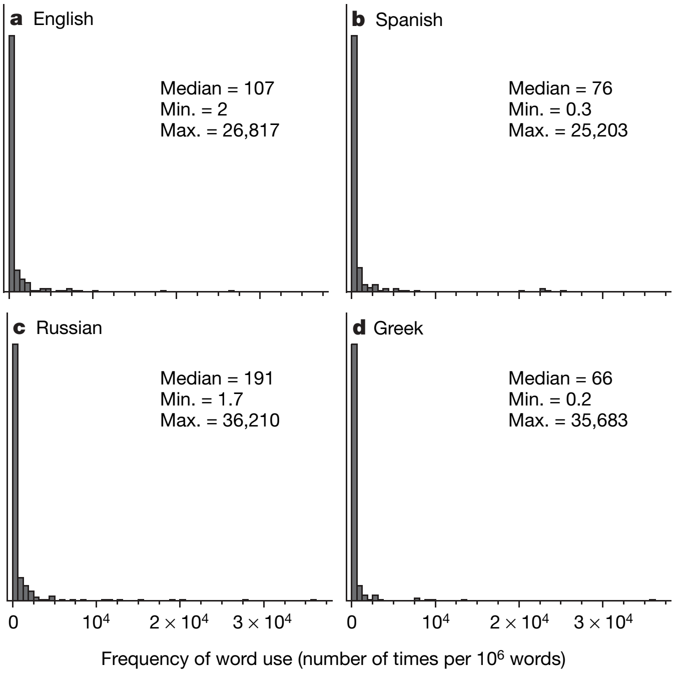

Figure 2 | Distribution of frequency of meaning-use for 200 meanings in four Indo-European languages. a, English; **b**, Spanish; **c**, Russian; and **d**, Greek. These four languages sample from across the Indo-European language family. Supplementary Fig. 1 shows where each language fits on the Indo-European phylogeny. Word-use frequencies are highly correlated among the four languages ($0.78 < r < 0.89$, mean $r = 0.84$; Supplementary Fig. 2).

Figure 3 plots the rate of lexical replacement against frequency of word-use for the 200 meanings. Separately in each corpus we observe a negative relationship (bold black line) between word frequency and rate (English, $r = -0.37$; Spanish, $r = -0.35$; Russian, $r = -0.41$; and Greek, $r = -0.32$: all $P < 0.0001$). In all four languages, the more a meaning is used today, the slower its rate of evolution has been throughout the 6,000- to 10,000-year history of Indo-European.

Some parts of speech are used more than others, so it is possible that the observed relationship arises from an effect of part of speech on rates of evolution. To examine this effect, we categorized meanings as either nouns, adjectives, verbs, pronouns, numbers, conjunctions, prepositions or special adverbs ('what', 'when', 'where', 'how', 'here', 'there' and 'not'). We then predicted variation in rates of lexical replacement from a regression model allowing both of these effects to operate simultaneously, thereby controlling for one another. The inverse relationship between frequency of meaning-use and rate of lexical evolution holds separately for parts of speech

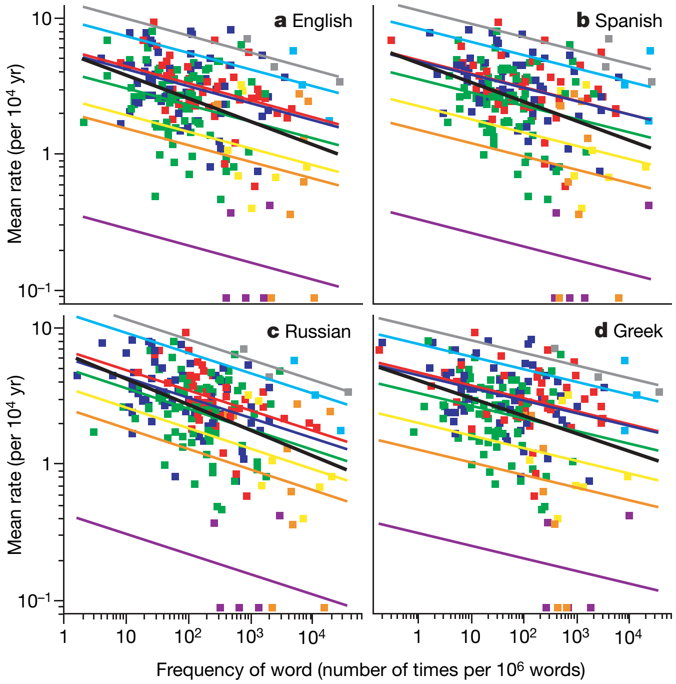

Figure 3 | Frequency of meaning-use plotted against estimated rate of lexical evolution for 200 basic meanings in four Indo-European languages. a, English; **b**, Spanish; **c**, Russian; and **d**, Greek. A linear regression (bold black line) reveals a consistent negative relationship between log(frequency of meaning-use) and log(rate of lexical replacement) across all four languages (English, $r = 0.37$; Spanish, $r = 0.35$; Russian, $r = 0.41$; and Greek, $r = 0.32$). Points are colour coded according to part of speech (see below). Coloured lines show the results of a multiple regression including frequency and part of speech. All relationships are negative (English, $R = 0.69$; Spanish, $R = 0.69$; Russian, $R = 0.71$; and Greek, $R = 0.69$). The height of each line above the x-axis indicates the relative speed of lexical evolution for a given frequency of meaning-use for each part of speech. Conjunctions (grey) evolve fastest, followed by prepositions (turquoise), adjectives (red), verbs (blue), nouns (green), special adverbs (yellow), pronouns (orange) and numbers (purple).

(Fig. 3, coloured lines). For a given frequency of meaning-use, prepositions and conjunctions evolve most quickly, followed by progressively slower evolution for adjectives, verbs, nouns, special adverbs, pronouns and finally numbers. The rank order of effects for part of speech is identical across the four corpora, and the combined models account for approximately 50% of the variance in rates of lexical evolution (English, $R = 0.69$; Spanish, $R = 0.69$; Russian, $R = 0.71$; and Greek, $R = 0.69$: all $P < 0.0001$; R denotes correlation derived from multiple regression). Adding an interaction effect between part of speech and frequency of meaning-use did not improve the fit of the model: frequency of meaning-use affects the rate of evolution in the same way for each part of speech.

The consistent pattern across the four languages in the relative rates of lexical replacement among parts of speech may help us to understand the mechanism by which word-use frequency affects rate of evolution. Frequency of word-use could directly modify the rate at which new word forms arise, with fewer spontaneous errors occurring for highly expressed words. Errors in word perception, recall and production have been shown to decrease with word frequency according to the 'power law of learning'[21,22]. Alternatively, the rate at which new forms appear could be the same for all meanings, with frequency of use affecting the probability that a population of speakers will come to adopt a given innovation. This suggests that some form of linguistic, frequency-dependent, purifying selection is responsible for the slow rate of evolution of highly expressed words. Innovations, being rare, may not be favoured in speech because there is an increased chance that they will be misinterpreted. The effect should be stronger the more often the meaning is required in speech or the more important it is to the meaning of speech.

In each of our language corpora, numbers, pronouns and the special adverbs evolve the most slowly for a given frequency of word-use. These parts of speech seem important to the meaning of spoken communication, and may therefore be subject to stronger selection. The rapidly evolving parts of speech include conjunctions, prepositions and adjectives whose exact forms may often be less important to conveying meaning. There is also evidence in natural populations of speakers that when more than one word is used to express the same meaning, the relative frequencies of use of the rare words is lower than expected from a neutral drift model of evolution[23], consistent with selection against innovations. These findings may then indicate that the purifying selection model of word evolution provides a more accurate description of how words evolve within populations of speakers. They are also consistent with models of cultural and linguistic evolution that incorporate a conformist bias[16,17], although these models cannot identify a priori which words will be subjected most strongly to such effects.

Our findings, based on a sample of fundamental vocabulary items, identify a general mechanism of linguistic evolution, which is expected to operate across all languages and timescales and makes predictions about rates associated with specific meanings. To the extent that the structure and everyday functions of human verbal communication mean that some words will tend to be used more frequently in all languages, we expect these words to evolve slowly, and vice versa for infrequently used words. Combined with parts of speech, this simple factor allows us to account for about 50% of the variance in rates of lexical replacement throughout the 6,000- to 10,000-year history of Indo-European languages. Given the many social, cultural and cognitive factors that can influence language[11–14], it is striking that word-use frequency alone can explain such a large proportion of the historical variation in rates of evolution. The generality of this influence is suggested in the finding that estimates of the rate of lexical replacement in Indo-European languages are correlated with rate estimates in Bantu[10], Cushitic and Malayo-Polynesian[1].

Being able to link variation in rates of lexical replacement to the frequency of word-use also provides insights into some features of comparative linguistics. One is that we expect languages to diverge initially in the least frequently used parts of their vocabularies. This may mean that languages retain mutual intelligibility far longer than expected from simple uniform rates models of linguistic divergence[8]. Within English, for example, words spoken at a higher frequency are more likely to be of Old English origin[24]. Related to this, the words for frequently used meanings should, on average, be less prone to borrowing during language contact. Higher frequency words may also be more likely to exhibit ancestral morphology. Irregular verbs in English often retain their ancestral morphology, and are among the most frequently expressed verbs[25]. Finally, we note that our rate estimates show that some words evolve slowly enough to allow homologous lexical forms to persist for tens of thousands of years. These slow rates demonstrate that humans are capable of producing a culturally transmitted replicator that, perhaps because of the purifying force of spoken word frequency, can have a replication accuracy as high as that of some genes[26]. Along with continued efforts at identifying cognate words separated by thousands of years of sound change[27], this raises the possibility of using selected lexical items to evaluate hypothesized 'long-range' linguistic relationships such as Eurasiatic[28] and Nostratic[29].

METHODS SUMMARY

Cognate data. We grouped the words for each of the 200 meanings in the fundamental vocabulary of ref. 8, based on previously published Indo-European lexical data[6] (Methods). Meanings had between 1 and 46 cognate sets across the 87 languages in our study, producing a total of 4,049 cognates (including unique word-forms).

Phylogenetic trees. We inferred a bayesian posterior distribution of phylogenetic trees of the 87 languages from a binary data matrix derived from the cognacy classifications[30]. The 4,049 binary vectors in the matrix code for the presence ('1') or absence ('0') of each of the 4,049 cognates. The consensus tree of this posterior sample is reported in Supplementary Fig. 1.

Rates of lexical replacement. We categorized words for each of the 200 meanings into k states representing the k different cognate classes identified for that meaning. For example, $k = 24$ for the meaning class 'big' because it has 24 cognates among the 87 languages. For each meaning, we estimate the instantaneous transition rate, q, from any state (cognate class) i to any state j, as the mean of its bayesian posterior distribution of rates, summed over models of evolution and phylogenetic trees in the posterior sample of trees. This accounts for uncertainty in the model of evolution and in the phylogenetic tree, and does not suffer from loss of information owing to the conversion of cognate data to pair-wise similarity scores between languages. Half-life estimates were derived as described in ref. 10.

Word frequency data. We obtained word-use frequencies from English[2], Spanish[3], Russian[4] and Greek[5] corpora, combining the frequencies for all words comprising a shared canonical form (for example, 'push', 'pushes', 'pushing' and 'pushed').

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 30 April; accepted 17 August 2007.

1. Kruskal, J. B., Dyen, I. & Black, P. D. in *Mathematics in the Archaeological and Historical Sciences* (eds Hodson, F. R., Kendall, D. G. & Tautu, P.) 361–380 (Edinburgh Univ. Press, Edinburgh, UK, 1971).
2. Leech, G., Rayson, P. & Wilson, A. *Word Frequencies in Written and Spoken English: based on the British National Corpus* (Longman, London, 2001).
3. Davies, M. Corpus del Español. ⟨http://www.corpusdelespanol.org⟩ (2001–02).
4. Sharoff, S. in *Corpus Linguistics Around the World* (eds Archer, D., Wilson, A. & Rayson, P.) 167–180 (Rodopi, Amsterdam, 2005).
5. Institute for Language and Speech Processing (ILSP) Corpus. Hellenic National Corpus (HNC) Web Version 3.0 [in Greek]. ⟨http://hnc.ilsp.gr/en/⟩ (1999–2006).
6. Gray, R. D. & Atkinson, Q. D. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426,** 435–439 (2003).
7. Gordon, R. G. *Ethnologue: Languages of the World* 15th edn (SIL International, Dallas, 2005).
8. Swadesh, M. Lexico-statistic dating of prehistoric ethnic contacts. *Proc. Am. Phil. Soc.* **96,** 453–463 (1952).
9. Dyen, I., James, A. T. & Cole, J. W. L. Language divergence and estimated word retention rate. *Language* **43,** 150–171 (1967).
10. Pagel, M. & Meade, A. in *Phylogenetic Methods and the Prehistory of Languages* (eds Clackson, J., Forster, P. & Renfrew, C.) 173–182 (MacDonald Institute for Archaeological Research, Cambridge, UK, 2006).

11. Labov, W. *Principles of Linguistic Change: Social Factors* (Blackwell, Oxford, UK, 2001).
12. Milroy, J. & Milroy, L. Linguistic change, social network and speaker innovation. *J. Linguist.* **21,** 229–284 (1985).
13. Nettle, D. Is the rate of linguistic change constant? *Lingua* **108,** 119–136 (1999).
14. Thomason, S. G. & Kaufman, T. *Language Contact, Creolization, and Genetic Linguistics* (Univ. California Press, Berkeley, 1988).
15. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, UK, 1983).
16. Boyd, R. & Richerson, P. J. *Culture and the Evolutionary Process* (Univ. Chicago Press, Chicago, 1985).
17. Kirby, S. *Function, Selection, and Innateness: the Emergence of Language Universals* (Oxford Univ. Press, Oxford, UK, 1999).
18. Croft, W. *Explaining Language Change: an Evolutionary Approach* (Longman, Harlow, UK, 2000).
19. Gimbutas, M. The beginning of the Bronze Age in Europe and the Indo-Europeans 3500–2500 B.C. *J. Indo-Eur. Stud.* **1,** 163–214 (1973).
20. Renfrew, C. *Archaeology and Language: the Puzzle of Indo-European Origins* (Cape, London, 1987).
21. Anderson, J. R. Acquisition of cognitive skill. *Psychol. Rev.* **89,** 369–406 (1982).
22. Ellis, N. C. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Stud. Second Lang. Acquisit.* **24,** 143–188 (2002).
23. Fontanari, J. F. & Perlovsky, L. I. Solvable null model for the distribution of word frequencies. *Phys. Rev. E* **70,** 042901 (2004).
24. Zipf, G. K. Prehistoric 'cultural strata' in the evolution of Germanic: The case of Gothic. *Mod. Lang. Notes* **62,** 522–530 (1947).
25. Francis, W. N., Kuçera, H. & Mackie, A. W. *Frequency Analysis of English Usage: Lexicon and Grammar* (Houghton Mifflin, Boston, 1982).
26. Burger, J., Kirchner, M., Bramanti, B., Haak, W. & Thomas, M. G. Absence of the lactase-persistence-associated allele in early Neolithic Europeans. *Proc. Natl Acad. Sci. USA* **104,** 3736–3741 (2007).
27. Mackay, W. & Kondrak, G. in *Proceedings of the 9th Conf. on Computational Natural Language Learning (CoNLL)* 40–47 (ACL, Schroudsburg, PA, 2005).
28. Greenberg, J. H. *Indo-European and its Closest Relatives: The Eurasiatic Language Family* Vol. 1, *Grammar* (Stanford Univ. Press, Stanford, CA, 2000).
29. Kaiser, M. & Shevoroshkin, V. Nostratic. *Annu. Rev. Anthropol.* **17,** 309–329 (1988).
30. Pagel, M. & Meade, A. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol.* **53,** 571–581 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. Gray and S. Greenhill for comments and advice. This research was supported by a grant to M.P. from the Leverhulme Trust.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.P. (m.pagel@reading.ac.uk).

METHODS

Cognate data. We used the comparative Indo-European database[31], which records word forms and cognacy judgements in 95 languages across the 200 terms in the fundamental vocabulary of ref. 8. We excluded 11 of the speech varieties that had not been coded in ref. 31 and were identified by those authors as less reliable, leaving 84 languages. We added cognacy judgements for the same 200 meanings for three extinct Indo-European languages (Hittite, Tocharian A and Tocharian B), based on multiple sources, for a combined sample of 87 languages[6]. Meanings had from between 1 and 46 cognate sets across the 87 languages, for a total of 4,049 cognates, including unique words.

Phylogenetic trees. We inferred the posterior distribution of phylogenetic trees for the 87 languages using the MCMC methods[32] implemented in *BayesPhylogenies*[30,33]. The cognacy data were transformed to a binary matrix, with rows representing the 87 languages and columns identifying the presence ('1') or absence ('0') of each of the 4,049 cognates (including unique 'cognates'). We then characterized the probability of these data on phylogenetic trees, using a two-state (presence/absence) continuous-time Markov transition rate model[10]. The probability of the data D given the model of evolution M and a tree T is written in the usual way as $P(D|M,T) = \Pi_c P(D|\mathbf{Q}_b,T)$, where here the model of evolution for the binary vectors $\mathbf{Q}_b$ is the 2×2 matrix recording the rates of transition between the binary elements corresponding to the gain of the cognate class and the loss of the cognate class, and the product is over the $c = 4{,}049$ binary vectors that identify cognate classes. The elements of $\mathbf{Q}_b$ are given by $q\pi_j$, where j represents the state (presence/absence) to which the cognate is moving and π is the equilibrium frequency of the jth state[10,34]. We used a single rate parameter q, and estimated the equilibrium frequencies of presence and absence from the data as part of the Markov chain. Transition rates were allowed to vary among cognate class vectors (successive vectors of the binary matrix) according to a gamma distribution[35] with four rate categories. The shape parameter of the gamma distribution was estimated from the data.

We derived the posterior sample of trees from a Markov chain allowed to run for 40,000,000 generations. After discarding the first 2,500,000 generations as burn-in, we sampled every 50,000th tree in the chain to ensure that successive trees were statistically independent. This produced a posterior sample of 750 trees. Examination of autocorrelation times of the MCMC plots indicated that runs had converged to the equilibrium distribution and showed very low autocorrelation, yielding an effective sample size of at least 500. The consensus tree of this posterior sample is reported in Supplementary Fig. 1.

Rates of lexical replacement. We categorized the words for each of the 200 meanings into k states representing the k different cognate classes identified for that meaning. In the meaning class 'big', for example, $k = 24$ because this meaning is represented by 24 cognates among the 87 languages. The probability of observing the distribution of the k lexical terms for a meaning m on any given tree can be written as $P(m|\mathbf{Q}_m,T)$ where $\mathbf{Q}_m$ is a $k \times k$ matrix of the transition rates from any cognate class i to any other class j for a particular meaning, and T is the phylogenetic tree[10]. The elements of $\mathbf{Q}_m$ are given by $q\pi_j$, where q is the instantaneous transition rate (as above but now for a particular meaning) and π_j is the equilibrium frequency of state j (ref. 10). We estimate $\mathbf{Q}_m$ using a continuous-time Markov transition rate model[10,30,33,34]. The equilibrium frequencies are not known and therefore must either be estimated from the data or fixed at prior values. For the results reported in Figs 1 and 3, we assumed uniform equilibrium frequencies across cognates. Thus for a meaning represented by $k = 3$ cognates, each π_j is set to 1/3. Because of the large number of cognate classes for some meanings ($k = 46$ for the meaning 'dirty'), it is impractical to estimate π_j from the data. However, using the observed empirical frequencies of the k classes for each meaning across the 87 languages gives the same qualitative results reported in Fig. 3, and none of our conclusions is altered.

The posterior distribution of the rate parameter q is estimated from a Markov chain that simultaneously proposes new values for q and samples new trees from the posterior distribution of T. We used the mean of the posterior distribution of q to estimate the rates of lexical replacement reported in Figs 1 and 3. Our approach accounts for uncertainty in the model of evolution and in the phylogenetic tree, and, unlike earlier lexicostatistical approaches to estimating rates of cognate replacement, does not suffer from information loss owing to the conversion of cognate data to pair-wise similarity scores between languages[36]. Mean and variance for the 200 meaning category rate estimates are provided in Supplementary Table 2. Rate estimates are not expected to be biased across parts of speech by the process of grammaticalization (see Supplementary Information for details).

Half-life estimates were calculated from the mean q values by solving $P = \mathrm{e}^{-qt}$ for t, setting $P = 0.5$ as described in ref. 10.

Word frequency data. Word-use frequencies were obtained for English, Spanish, Russian and Greek from the corpora databases described in refs 2–5. Word frequencies were compiled by searching for all forms listed under the canonical form (or lemma) of each meaning. For example, for the verb meaning 'push' in English, we include 'push', 'pushes', 'pushing' and 'pushed'. Word frequency and part of speech data are provided in Supplementary Table 2.

31. Dyen, I., Kruskal, J. B. & Black, P. An Indo-European classification, a lexicostatistical experiment. 1. *Trans. Am. Phil. Soc.* **82,** 1–132 (1992).
32. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21,** 1087–1091 (1953).
33. Pagel, M. & Meade, A. in *Mathematics of Evolution and Phylogeny* (ed. Gascuel, O.) 121–139 (Oxford Univ. Press, New York, 2005).
34. Pagel, M. & Meade, A. in *The Evolution of Cultural Diversity: a Phylogenetic Approach* (eds Mace, R., Holden, C. J. & Shennan, S.) 235–256 (UCL Press, London, 2005).
35. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39,** 306–314 (1994).
36. Steel, M. A., Hendy, M. D. & Penny, D. Loss of information in genetic distances. *Nature* **336,** 118 (1988).