PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2011) **366**, 1101–1107
doi:10.1098/rstb.2010.0315

Research

How do we use language? Shared patterns in the frequency of word use across 17 world languages

Andreea S. Calude and Mark Pagel*

School of Biological Sciences, University of Reading, Reading, UK

We present data from 17 languages on the frequency with which a common set of words is used in everyday language. The languages are drawn from six language families representing 65 per cent of the world's 7000 languages. Our data were collected from linguistic corpora that record frequencies of use for the 200 meanings in the widely used Swadesh fundamental vocabulary. Our interest is to assess evidence for shared patterns of language use around the world, and for the relationship of language use to rates of lexical replacement, defined as the replacement of a word by a new unrelated or non-cognate word. Frequencies of use for words in the Swadesh list range from just a few per million words of speech to 191 000 or more. The average inter-correlation among languages in the frequency of use across the 200 words is 0.73 ($p < 0.0001$). The first principal component of these data accounts for 70 per cent of the variance in frequency of use. Elsewhere, we have shown that frequently used words in the Indo-European languages tend to be more conserved, and that this relationship holds separately for different parts of speech. A regression model combining the principal factor loadings derived from the worldwide sample along with their part of speech predicts 46 per cent of the variance in the rates of lexical replacement in the Indo-European languages. This suggests that Indo-European lexical replacement rates might be broadly representative of worldwide rates of change. Evidence for this speculation comes from using the same factor loadings and part-of-speech categories to predict a word's position in a list of 110 words ranked from slowest to most rapidly evolving among 14 of the world's language families. This regression model accounts for 30 per cent of the variance. Our results point to a remarkable regularity in the way that human speakers use language, and hint that the words for a shared set of meanings have been slowly evolving and others more rapidly evolving throughout human history.

Keywords: language evolution; frequency of use; word evolution

1. INTRODUCTION

There is now a growing feeling among researchers that elements of human language can be studied as discrete entities that are transmitted from mind to mind and evolve by a process of descent with modification [1]. Languages can be transmitted with a surprising degree of fidelity, and the many parallels between linguistic and genetic evolution mean that approaches drawn from the fields of phylogenetics and comparative biology are increasingly being applied to study languages. Phylogenies of languages chart the history and movement of human cultures [2–5], and elements of language can be studied to understand the social, cultural and linguistic factors that govern their rates and patterns of change through time [1]. Our interest here is to examine the generality of one force known from previous work [6] to influence rates of lexical evolution, that being the frequency with which words are used in everyday speech.

If words are thought of as one of the discrete units of a language, they show what molecular geneticists would refer to as rate heterogeneity, with some evolving at high rates and others at far slower rates. For example, among a sample of 87 Indo-European languages, all speakers use a related group of sounds or words to describe 'two' (we use the symbol <'> to denote a given meaning, or concept, and the symbol <"> to refer to a word form) objects but use 45 or more different and unrelated words to describe something as 'dirty' [6,7]. The related sounds for the word "two" are all homologues or what linguists would refer to as cognates—words that derive by descent with modification from a common ancestral word. The 45 different ways of expressing the idea of 'dirty' thus represent at least 45 newly produced or non-cognate words in the 9000 or so years since the Indo-European languages descended from their common proto-language. The rate at which new non-cognate words arise can be studied phylogenetically using language phylogenies and appropriate statistical models [1,6,8]. Applied to a sample of Indo-European language trees, we have found that the quantitative rates of change for "two" and "dirty" differ about 100-fold [6].

Why do the words for some meanings evolve so rapidly and others slowly? In a previous report [6], we described a general evolutionary law relating the

* Author for correspondence (m.pagel@reading.ac.uk).

One contribution of 26 to a Discussion Meeting Issue 'Culture evolves'.

frequency with which words are used in everyday speech to rates of lexical replacement, defined as the replacement of a word by a new unrelated or non-cognate word. Measured across the Indo-European languages, frequently used words have slower rates of lexical replacement than infrequently used words. We reached this conclusion from studying linguistic corpora for four phylogenetically widely spaced Indo-European languages: Greek, Russian, Spanish and English. Linguistic corpora record, among other things, the frequencies with which speakers use a wide range of words in their everyday speech (tables 1 and 2). Greek is a basal member of the Indo-European language tree, Russian is part of the Slavic language family, Spanish is one of the Romance languages and English is a Germanic language.

We studied the frequencies of use in each of these languages for the 200 words that make up the Swadesh fundamental vocabulary word list [10]. The list comprises 200 common meanings, such as 'mother', 'lake', 'mountain', 'three', 'red', 'green', 'to vomit', 'to kill', 'dirty' and 'dull', that Swadesh thought would be present in all languages, much like one might expect there to be a universal set of genes among biological organisms. The list avoids technical terms and specific environmental terms. It would be possible to construct a different list, but the Swadesh list has formed the principal basis for pursuing historical reconstructions and for investigating language history for the past 60 years. The list is commonly used to infer linguistic phylogenies, and it is the set of words that we used to measure rates of lexical replacement in our earlier work.

Despite being separated by thousands of years of linguistic evolution, the average inter-correlation among the four languages in the frequency with which they used these common words was 0.85. This very high average inter-correlation suggests that speakers of different languages use language in the same way and probably for the same purpose. The phylogenetic placement of the four languages we studied further suggests that frequency of use is a stable trait, leading to the speculation that the frequencies we observe in these extant languages are representative of the ancestral or proto-Indo-European languages. If word-use frequencies are a stable and fundamental feature of human language use in general, this leads to the intriguing possibility that the words for a shared set of meanings will be slowly evolving and others more rapidly evolving in all of the world's languages, and that this will probably have been true throughout human history. This is to say that both the frequencies of use and the rates of lexical replacement we found for the Indo-European languages might be representative more broadly of human language evolution.

Pagel [1] reports some evidence in support of this speculation. Figure 1 (re-drawn from [1]) plots the rates of lexical replacement for the Indo-European languages [6] against a list of 110 words that the late Russian comparative linguist Sergei Starostin identified as among the most stable in 14 language families from around the world [11]. Starostin's list is a subjective rank-ordering based on his work with these language families from the most stable (rank = 1) or slowly evolving to less stable (rank =110). The figure

Table 1. Language corpora consulted by language family. The corpus size is given from the documentation of each of the corpora used. The language classification is given from the online Ethnologue database [9].

language family	language	size (no. of words)
Indo-European	English	100 million
	Russian	140 million
	Greek	47 million
	Portuguese	45 million
	Spanish	1 million
	Chilean Spanish	450 million
	French	31 390 000
	Czech	100 million
	Polish	450 million
Sino-Tibetan	Chinese	1 million
Uralic	Finnish	21 329 990
	Estonian	1 million
Niger-Congo	Swahili	2 million
Altaic	Turkish	2 million
Austronesian	Māori	1 million
unclassified languages	Basque	5 million
Creole	Tok Pisin	864 900

Table 2. Sources of the corpora.

Basque	twentieth Century Corpus of Basque, http://www.uzei.com/
Chilean Spanish	Scott Sadowsky, LIFCACH, http://www2.udec.cl/~ssadowsky
Chinese (Mandarin)	Lancaster Corpus of Mandarin Chinese, http://corpus.leeds.ac.uk
Czech	Czech National Corpus, http://ucnk.ff.cuni.cz/english/kdejsme.php
English	BNC, http://www.natcorp.ox.ac.uk
Estonian	Corpus of Written Estonian, http://www.cl.ut.ee/korpused
Finnish	Parole Corpus of Finnish, http://kaino.kotus.fi/sanat/taajuuslista/parole_5000.html
French	Frantext, http://www.atilf.fr/frantext.htm
Greek	HNC, http://hnc.ilsp.gr/en
Māori	Māori Broadcasting Corpus, Boyce, M. T. 2006 A corpus of modern spoken Māori. Unpublished PhD thesis available in the library at Victoria University of Wellington.
Polish	Polish National Corpus, http://nkjp.pl
Portuguese	Mark Davies, http://www.corpusdoportugues.org/x.asp
Russian	Sharoff, S. Corpus linguistics around the world (eds Archer, D., Wilson, A. & Rayson, P.), pp. 167–180 (Rodopi, Amsterdam, 2005), http://www.ruscorpora.ru/
Spanish	Mark Davies, http://www.corpusdelespanol.org
Swahili	Helsinki Corpus of Swahili, http://www.aakkl.helsinki.fi/cameel/corpus
Tok Pisin	Slone Wantok Corpus, http://www.tokpisin.org
Turkish	METU Turkish Corpus, http://www.ii.metu.edu.tr/corpus

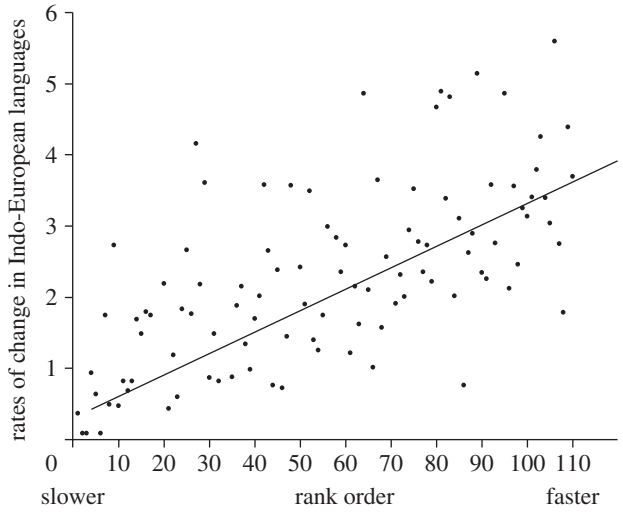

Figure 1. Re-drawn from Pagel [1]. Statistically estimated rates of lexical replacement for 110 words in the Swadesh list in the Indo-European languages (from [6]) correlated with rank-ordering of subjectively assessed rates of change for the same words in a worldwide sample of 14 language families [11]. The correlation $r = 0.65$. The language families include Sino-Tibetan, Austroasiatic, Altaic, Austronesian, Australian, Khoisan, North Caucasian, Dravidian, Indo-European, Kartvelian, Afroasiatic, Tai, Uralic and Yenisan.

shows that slowly evolving words in Indo-European languages are also slowly evolving in the world's other language families, and vice versa: rates of evolution might indeed have been conserved throughout human history.

Here, we wish to examine these ideas further by collecting data on frequencies of word use from languages around the world, and relating those frequencies to rates of lexical replacement and to Starostin's list.

2. DATA AND METHODS

We collected data on the frequency of word use for the 200 Swadesh word list items from linguistic corpora describing 17 languages (table 1). The languages derive from six language families (Austronesian, Altaic, Indo-European, Niger-Congo, Sino-Tibetan and Uralic), plus one unclassified language (Basque), and a creole language (Tok Pisin). The families are widely geographically spaced and represent 65 per cent of the world's 7000 or so extant languages [9]. The corpora range in size from one million recorded words (Chinese, Estonian, Māori and Spanish) to 450 million words (Chilean Spanish and Polish; table 1). The corpora include spoken and written language use from a variety of genres, including spontaneous conversation, academic writing, newspaper articles and radio transcripts. The Tok Pisin corpus is smaller and less balanced than the others, but we include it here for its interest as a creole.

We normalized all frequency-of-use data from table 1 to a common basis of frequency of use per one million words. The Indo-European languages are disproportionately represented, so we calculated a mean Indo-European frequency-of-use score for the nine Indo-European languages (treating Chilean Spanish as Indo-European). There were 70 words out of the 17 languages $\times$ 200 Swadesh list items, or 2 per cent of the total, for which frequency data were not available. We replaced these missing data with the mean frequency calculated from the other languages and again using the mean Indo-European frequency rather than the separate Indo-European data points. If a word was missing from one of the Indo-European languages, we used the others to calculate the IE mean.

We added to these frequency data, information from our previous work [6] on the rates of lexical replacement in the Indo-European languages for each of the meanings in the Swadesh word list. These rates were estimated using a statistical likelihood model of word evolution [7] applied to phylogenetic trees derived from 87 Indo-European languages. The number of cognate classes (the number of distinct unrelated sets of words) for a given meaning varied from 1 (e.g. 'two') to 46 (e.g. 'dirty'). For each of the 200 meanings, we calculated the mean of the posterior distribution of rates as derived from a Bayesian Markov chain Monte Carlo model that simultaneously accounts for uncertainty in the parameters of the model of cognate replacement and in the phylogenetic tree of the languages. Rate estimates were scaled to represent the expected number of cognate replacements per 10 000 years, assuming an 8700 year age for the Indo-European language family [2]. We used these Indo-European rates because they are as yet the only published rates based on statistical modelling applied to phylogenies.

The Indo-European rates of lexical replacement vary roughly 100-fold. At the slow end of the distribution, the rates predict 0–1 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 fundamental vocabulary of 1–46 distinct cognate classes among the different meanings. These rates can be converted to estimates of the linguistic half-life [6,12], or the time in which there is a 50 per cent 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.

3. RESULTS

(a) *Frequency of use*

We logarithmically transformed the frequency data prior to analyses. The average inter-correlation among the languages in the frequencies of use across the 200 word meanings is 0.73 ($p < 0.0001$), using the single Indo-European mean. Previously, we found an average inter-correlation of 0.85 for English, Russian, Greek and Spanish [6], and here we find an average inter-correlation among the nine Indo-European languages of 0.82. To summarize these correlations, we derived the first principal component

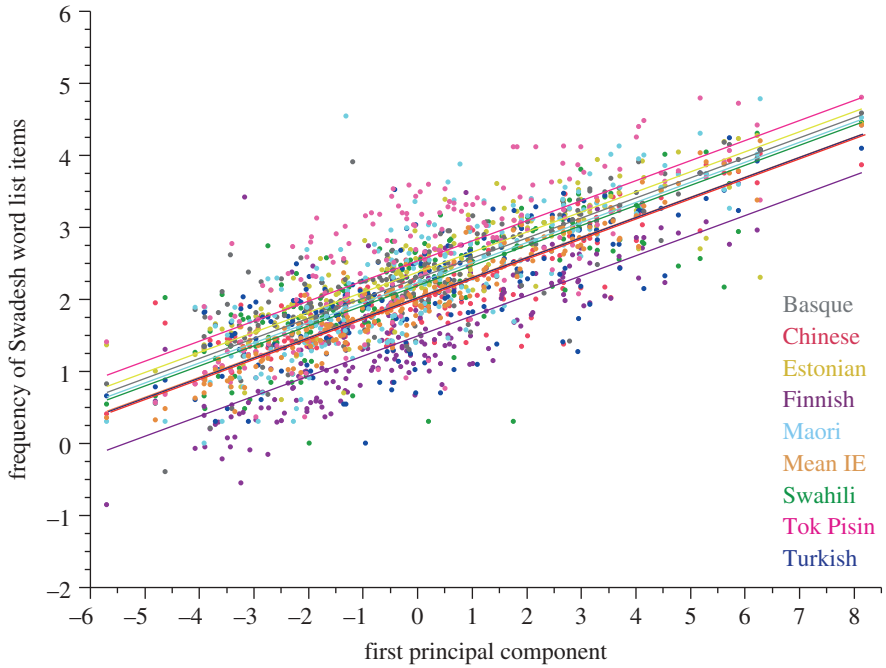

Figure 2. Log-transformed frequency of use per million words for each of the eight languages plus the Indo-European mean vector of nine languages (see table 1 and text) plotted against the first principal component factor scores of frequencies. The first principal component uses the mean IE-vector together with the other eight languages so as not to bias towards the Indo-European (see text). We fitted regression lines to each language, allowing different intercepts but constraining lines to be parallel simply for illustrative purposes. Fitted this way, the overall relationship accounts for 69.4% of the variance in the principal component. The positive slopes indicate that each language's frequencies of use correlate positively with the principal component that summarizes them. Allowing slopes to vary increases the percentage to 70.3% (not significantly different).

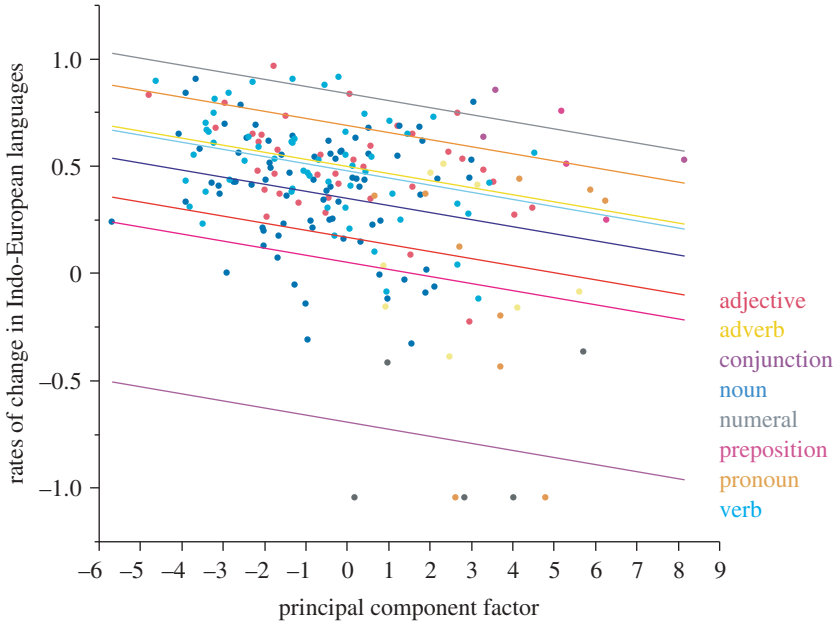

Figure 3. Statistically estimated rates of lexical replacement for the Swadesh list in the Indo-European languages (from [6]) correlated with the first principal component loadings obtained from the frequency of use of the Swadesh list among 17 languages, and with part of speech. The relationship accounts for 46% of the variance in rates of lexical replacement ($p < 0.0001$), and holds separately within part-of-speech category: prepositions ('in', 'with'), conjunctions ('and', 'because'), adjectives ('white', 'thin'), verbs ('to throw', 'to eat'), nouns ('hand', 'hair'), special adverbs ('here', 'some'), pronouns ('I', 'they') and numerals ('one', 'five'). Regression lines were allowed to have separate intercepts but constrained to have the same slope. Allowing these slopes to vary increased the overall R^2 to 0.47, not a significant change.

of the frequency data again using a single vector of the mean frequencies for the nine Indo-European languages. The first principal component was the only principal factor with an eigenvalue greater than 1.0 and accounts for 70.4 per cent of the variance. This figure includes several large outliers with plausible explanations (see discussion below as to what these might be) and so is probably conservative.

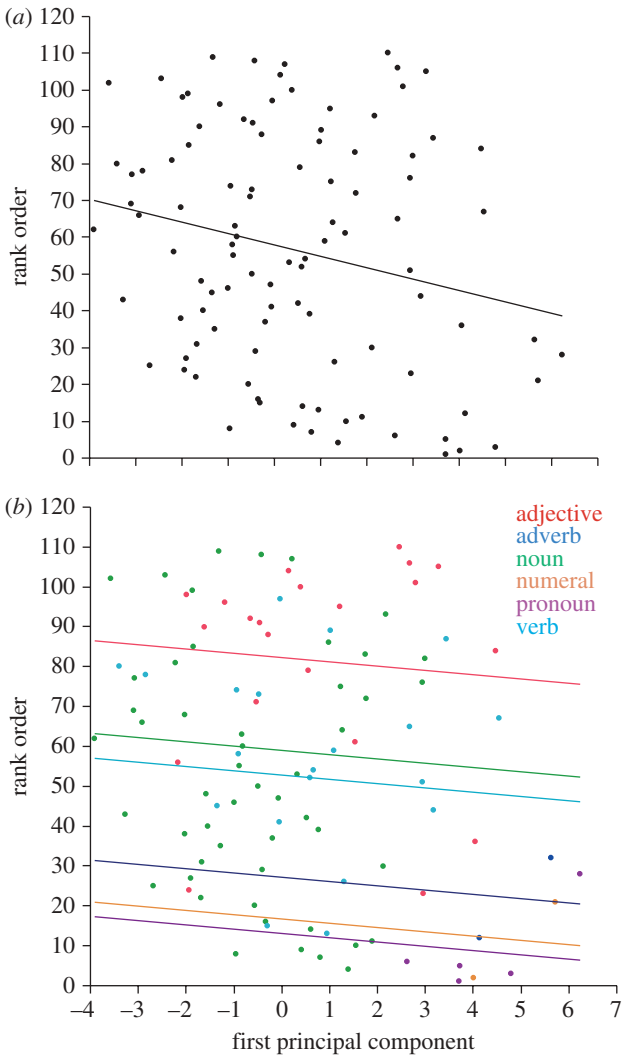

Figure 4. (*a*) The first principal component of frequency of use predicts the subjective rank-order rate of change for 110 words as judged in a worldwide sample of 14 language families [11], correlation $r = 0.22$, $p < 0.0263$. The language families are described in figure 1. (*b*) Regression model including the first principal component loadings and parts of speech $R^2 = 0.31$, $p < 0.0001$). Regression lines were allowed to have separate intercepts but were constrained to have the same slopes. Allowing separate slopes did not significantly increase the overall R^2.

The individual languages each fit the first principal component (figure 2) as we would expect from their high average inter-correlations. The different 'elevations' or *y*-axis intercepts of the languages are statistically different and might be of interest, but we cannot know whether they are artefacts of the reported size of each corpus. A corpus might report being based on 45 million utterances, but we cannot independently verify this. However, these mean differences do not influence correlations or the principal component. Where there are outliers on the plot, they are often specific to a particular language rather than to a set of languages and therefore probably arise from idiosyncratic language-specific factors. For example, the word "rotten" is used at a relatively high frequency in Finnish, but not in the other Uralic languages. The Finnish corpus is drawn principally from newspaper and magazine texts, and literature. Because much of Finland is low lying and makes contact with the Baltic, the Finnish corpus team suggested that many articles in the Finnish media focus on the consequent problems of rotting wood and damage caused to housing because of dampness. Similarly, in Māori, the word "ngā" meaning 'to breathe', is also used as a noun meaning 'breath' but it occurs in expressions such as "ngā … nā" ('those near you'), "ngā … nei" ('those near me') and "ngā … rā" ('those away from both speaker and listener'), and even as a definite plural article, meaning 'the'. The English verb "to know" is distributed across two finer grained distinctions in French, namely, "connaître" ('to know a person') and "savoir" ('to know a thing/fact/theory'). The word "louse" might have been used at a relatively high frequency by our hunter–gatherer ancestors, but now its frequency varies considerably among languages.

Other outliers might arise from issues of how to code some words. The Swadesh list item "day" refers to the period of daytime as opposed to the period of darkness that English speakers at least call "night". But languages including English, German and Māori also use the form meaning 'day' in the common greeting "Good day!". This formulaic use greatly increases the frequency of "day" in any corpus containing conversational data, or any dialogue (whether actual or fictional). In contrast, in Chinese, there are three choices for 'day': the formal version, "日" (which also means 'Japan', 'date' and 'sun'), the more informal character "天" (but this can also be used to mean 'sky', 'heavens', 'God', 'weather', 'nature', 'season') or the form "白天" (which actually means 'daytime'). The latter fits best the Swadesh word meaning intended; however, this is going to be much less frequently used in Chinese, in comparison to its cross-linguistics given that 'Good day!' in Chinese involves the "天" character, and not "白天".

(b) *Rates of lexical replacement*

If frequencies of use are a shared feature of human language and if frequencies predict rates of lexical replacement, then the principal component of frequencies from the worldwide sample should predict the rates of lexical replacement for the Indo-European languages. We predicted Indo-European rates from first factor loadings in a two-factor linear regression model including parts of speech coded as discrete categories. As expected, higher principal factor loadings are associated with lower rates of lexical replacement, and this relationship holds separately within parts of speech ($R^2 = 0.46$, $p < 0.0001$, figure 3). This result is comparable to the percentage of variance in rates of lexical replacement we were able to account for using the Greek, Spanish, Russian and English frequencies in our earlier study [6].

We repeated this analysis using a different principal component calculated from a dataset from which we had deleted the Indo-European languages. This removes any possibility of a correlation arising between the rates of replacement and frequency of use that might be true only of the Indo-European languages. This new principal component accounted for 67 per cent of the variance and returned an R^2 value in the multiple regression of 0.46 per cent ($p < 0.0001$), unchanged from the previous analysis.

(c) *Rank-order rates of change from a worldwide sample*

We repeated the regression model above, this time predicting Starostin's rank-order subjective ratings of stability for 110 words from the Swadesh word list. The first principal component is a significant predictor of rank order, and the overall model accounts for 30 per cent of the variance ($R = 0.54$, $p < 0.001$). Repeating this analysis using the modified principal component from which the Indo-European languages had been removed also returns an R^2 of 0.31 ($p < 0.0001$, figure 4).

4. DISCUSSION

Our results confirm our earlier speculation that the frequency with which a common set of words is used in everyday speech is a shared feature of human languages: to a reasonable first approximation, this appears to suggest that all human groups use language in a similar way, and probably for the same purpose. Pagel [13] has argued elsewhere that human language evolved to allow people to vary how they are perceived in the social phenotype of human culture in a manner analogous to the ways that genes use gene regulation to vary their expression in organismal phenotypes. In both cases, a form of digital communication—language or gene regulation—is used to influence how a replicating entity is exposed to the outside world. Unlike all other animal societies, human culture is based on elaborate specialization, exchange and division of labour among unrelated people. These complex reciprocal relationships are inherently laced with commonalities and conflicts of interest because everyone in a human society is free to pursue their own reproductive interests.

Language is the means by which we achieve a precise and nuanced communication system to manage how we are seen by others, and to influence how others are seen. Language permits people to enhance their own contributions to relationships or exchanges, and perhaps gently to denigrate those of others, and more generally to keep track of who did what to whom, at what time and how often. Our cooperative societies depend on language to transmit this information about others' reputations as a way of promoting exchanges among unrelated people. Frequently used words in the Swadesh list include the pronouns and number words and the so-called special adverbs or "who", "what", "where", "why" and "when". The shared high frequency of use of these socially relevant words is consistent with this idea of language as a device for social regulation.

No one, of course, doubts that language is for communicating. Its value in transmitting knowledge, making plans and in teaching is obvious. But from a gene's eye view, communication is only valuable insofar as it influences another animal's behaviour in a way that serves the communicator. One problem with thinking of language as merely a system for transmitting information is that much of what we might share with someone else could benefit them, without returning any benefit to ourselves, or worse it might disadvantage us. If someone reveals where their favourite source of water is, they might then find themselves having to compete for that resource. This tells us to look for clues in the nature and use of language that point to how it benefits the speaker. Transmitting information can benefit speakers, but this benefit might often have less to do with the information itself than the cooperative or reciprocal relationship that an act of potential altruism encourages. On this view, being in a position to share information is an act with the potential to enhance one's value and prestige in other's eyes.

We find it remarkable that frequency of use and a word's part of speech can together account for close to half of the variation in rates of lexical replacement. The results using Starostin's rank-order list are encouraging that this might be a very general effect, and we look forward to testing whether our results predicting rates of lexical replacement hold in new samples using rates derived from other language families. Frequency of use might affect rates of lexical replacement by altering 'production errors'—akin to the mutation rate in genetics—or by altering the rate at which a new form is adopted in a speech community (akin to selection) or both [6,14]. Word use may be under strong purifying selection within populations of speakers, if only through the rule 'speak as most others do'. It is difficult to understand how entire populations of speakers could otherwise agree on a single or a small number of mostly arbitrary sounds to represent a given meaning. Such a rule would have been advantageous in our history if speakers who make mistakes are disadvantaged.

Some words may acquire connections in the cognitive or semantic space [15], connections the strength or size of which may influence how rapidly words evolve. For example, the Old English "gebed" meaning "prayer" (from the Old Proto-Indo-European root "*gwedh") became shortened to the form "bede" meaning "prayer bead" or rosaries used for prayer, from which we now have the modern English word "bead" (used widely in any necklaces and other cultural artefacts). This may suggest a third route by which frequency effects operate, that being to increase the chance that a word acquires connections to other words or meanings by virtue of being used in a variety of settings and situations.

Linguists are well aware that linguistic behaviour, sociolinguistic variation and language change are moderated to a great extent by the frequency with which words are used [14,16,17], but these studies have not investigated the link between frequency and rates of lexical replacement over periods of thousands of years. Language evolution and change are highly sensitive to frequency of use [6,18,19]. Frequency effects begin to play a role in building up of linguistic categories and sequential patterns right from the language development stage, as children acquire language through repetition [20], and continue through adulthood (adults are good at estimating the frequency of words in a given list, cf. [21]), as well as through the process of learning a second/foreign language [14]. High-frequency items behave differently and possess different characteristics from low-frequency items across all linguistic levels, from the graphic symbols used to write down texts, to the sound patterns involved in uttering them and the morphemes used to make up words, and including the grammatical structures observed [22].

5. CONCLUDING REMARKS

Our results point to a surprising regularity in the way that human speakers use language. It might be that the way we use language and its structure means that some words inevitably will be used more than others. If so, then this leads to the intriguing possibility that the words for a shared set of meanings will be slowly evolving and others more rapidly evolving in all of the world's languages, and that this will probably have been true throughout human history.

Other elements of language and culture might also be studied to try to understand the factors that influence their rates of change. An obvious next step for studying how frequency of use affects lexical replacement is to move 'down' one level to phonemes. Do these building-block sounds get replaced within words as part of the normal progress of lexical change, and is their rate of replacement influenced by how often they are used in language as a whole? Moving outside of language, what factors influence the rates of change of technological innovations or styles of fashion and art? It is less clear how to apply the idea of frequency of use in these examples, but there might be analogues. For example, how often a piece of technology is used, its contribution to a society's wealth or how widely adopted it is, might be related to the rate at which it evolves or adapts to societies' changing needs or whims.

We thank the Leverhulme Trust (M.P.) and the New Zealand Foundation for Research, Science and Technology (A.S.C.) for supporting this work. We are grateful to Heiki-Jaan Kaalep, Bilge Say, Scott Sadowsky, Arvi Hurskainen, Michal Kren, Piotr Pezik, Katherine Cao and Miriam Urkia for help with the corpus data. Chris Venditti and Andrew Meade helped with analyses.

REFERENCES

1 Pagel, M. 2009 Human language as a culturally transmitted replicator. *Nat. Rev. Genet.* **10**, 405–415.
2 Gray, R. D. & Atkinson, Q. D. 2003 Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426**, 435–439. (doi:10.1038/nature02029)
3 Gray, R., Drummond, A. & Greenhill, S. 2009 Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323**, 479–483. (doi:10.1126/science.1166858)
4 Holden, C. 2002 Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. *Proc. R. Soc. Lond. B* **269**, 793–799. (doi:10.1098/rspb.2002.1955)
5 Kitchen, A., Ehret, C., Assefa, S. & Mulligan, C. 2009 Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. *Proc. R Soc. B* **276**, 2703–2710. (doi:10.1098/rspb.2009.0408)
6 Pagel, M., Atkinson, Q. & Meade, A. 2007 Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* **449**, 717–720. (doi:10.1038/nature06176)
7 Pagel, M. & Meade, A. 2006 Estimating rates of lexical replacement on phylogenetic trees of languages. In *Phylogenetic methods and the prehistory of languages* (eds P. Forster & C. Renfrew), McDonald Institute Monographs, pp. 173–182. Cambridge, UK: McDonald Institute of Archaeology.
8 Pagel, M. 2000 Maximum likelihood models for glotto-chronology and for reconstructing linguistic phylogenies. In *Time-depth in historical linguistics* (eds C. Renfrew, A. MacMahon & L. Trask), pp. 189–207. Cambridge, UK: McDonald Institute of Archaeology.
9 Lewis, M. P. (ed.) 2009 *Ethnologue: languages of the world*, 16th edn. Dallas, TX: SIL International.
10 Swadesh, M. 1952 Lexicostatistic dating of prehistoric ethnic contacts. *Proc. Am. Phil. Soc.* **96**, 452–463.
11 Starostin, S. A. 2007 Languages of the Slavic culture. In *Works on linguistics* (ed. S. A. Starostin), pp. 827–839. Moscow: Nauka.
12 Pagel, M. 2000 The history, rate, and pattern of world linguistic evolution. In *The evolutionary emergence of language* (eds C. Knight, M. Studdert-Kennedy & J. Hurford), pp. 391–416. Cambridge, UK: Cambridge University Press.
13 Pagel, M. 2008 Rise of the digital machine. *Nature* **452**, 699. (doi:10.1038/452699a)
14 Ellis, N. 2002 Frequency effects in language processing: a review with implications for theories of implicit and explicit language acquisition. *Stud. Sec. Lang. Acquis.* **24**, 143–188.
15 Huettig, F., Quinlan, P. T., McDonald, S. A. & Altmann, G. T. M. 2006 Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. *Acta Psychol.* **121**, 65–80. (doi:10.1016/j.actpsy.2005.06.002)
16 Bybee, J. 2007 *Frequency of use and the organisation of language*. Oxford, UK: Oxford University Press.
17 Bybee, J. & Hopper, P. (eds) 2001 *Frequency and the emergence of linguistic structure*. Amsterdam, The Netherlands: Benjamins.
18 Croft, W. 2000 *Explaining language change: an evolutionary approach*. London, UK: Longman.
19 Kemmer, S. & Israel, M. 1994 Variation and the usage-based model. In *Papers from the thirtieth regional meeting of the Chicago Linguistics Society: Para-session on variation and linguistic theory*, vol. 2 (eds K. Beals, J. Denton, R. Knippen, L. Melnar, H. Suzuki & E. Zeinfeld), pp. 165–179. Chicago, IL: Chicago Linguistics Society.
20 Tomasello, M. 2003 *Constructing a language*. Cambridge, MA: Harvard University Press.
21 Shapiro, B. J. 1969 The subjective estimate of relative word frequency. *J. Verb. Learn. Verb. Behav.* **8**, 248–251. (doi:10.1016/S0022-5371(69)80070-8)
22 Bybee, J. & Thompson, S. 2000 Three frequency effects in syntax. *Berkeley Linguist. Soc.* **23**, 65–85.