

The deep history of the number words

Article

Accepted Version

Pagel, M. ORCID: <https://orcid.org/0000-0001-7287-8865> and Meade, A. ORCID: <https://orcid.org/0000-0001-7095-7711> (2018) The deep history of the number words. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 373 (1740). 0517. ISSN 0962-8436 doi: <https://doi.org/10.1098/rstb.2016.0517> Available at <https://centaur.reading.ac.uk/74672/>

It is advisable to refer to the publisher's version if you intend to cite from the work. See [Guidance on citing](#).

To link to this article DOI: <http://dx.doi.org/10.1098/rstb.2016.0517>

Publisher: The Royal Society

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

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online



PHILOSOPHICAL TRANSACTIONS B

The Deep History of the Number Words

Journal:	<i>Philosophical Transactions B</i>
Manuscript ID	RSTB-2016-0517.R1
Article Type:	Research
Date Submitted by the Author:	19-May-2017
Complete List of Authors:	Pagel, Mark; University of Reading, School of Biological Sciences Meade, Andrew; University of Reading
Issue Code: Click here to find the code for your issue.:	NUMERICAL
Subject:	Evolution < BIOLOGY
Keywords:	numerosity, word evolution, lexical replacement, rates of change, phylogenetics, linguistics

SCHOLARONE™
Manuscripts

The Deep History of the Number Words

Mark Pagel^{1,2} and Andrew Meade¹

1. School of Biological Sciences, University of Reading, Reading RG6 6UR, UK

2. The Santa Fe Institute, 1399 Hyde Park Rd, Santa Fe, NM 87501, USA

Corresponding author: Mark Pagel, m.pagel@reading.ac.uk

Keywords: numerosity, word evolution, lexical replacement, rates of change, phonetic, linguistics

For Review Only

Summary

We have previously shown that the 'low limit' number words (from one to five) have exceptionally slow rates of lexical replacement when measured across the Indo-European languages. Here we replicate this finding within the Bantu and Austronesian language families, and with new data for the Indo-European languages. Number words can remain stable for 10,000 to over 100,000 years, or around 3.5 to 20 times longer than average rates of lexical replacement among the Swadesh list of 'fundamental vocabulary' items. Ordinal evidence suggests that number words also have slow rates of lexical replacement in the Pama-Nyungan language family of Australia. We offer three hypotheses to explain these slow rates of replacement: i) that the abstract linguistic-symbolic processing of 'number' links to evolutionarily conserved brain regions associated with numerosity; ii) that number words are unambiguous and therefore have lower 'mutation rates'; and iii) that the number words occupy a region of the phonetic space that is relatively full and therefore resist change because alternatives are unlikely to be as 'good' as the original word.

1. Introduction

In previous work we introduced the formal study of rates of lexical replacement as estimated from statistical models applied to phylogenetic trees of languages [1]. By 'lexical replacement' we refer to the replacement over evolutionary time of a word for a given meaning by a new and non-cognate word. For example, the word *hand* in English is cognate to the German *hand* but not to the Spanish *mano*, which in turn derives from the Latin *manus*. Both the Germanic and Romance languages independently trace their ancestry back to a proto-Latin language. This suggests that the word *hand* is a newer and non-cognate form that probably arose somewhere along the lineage that eventually gave rise to the Germanic languages.

In our earlier study we found that rates of lexical replacement varied around 100-fold among the 200 items in the widely-used Swadesh fundamental vocabulary [2]. The Swadesh list includes words that might be expected to be found in all languages, such as common nouns, verbs, adjectives and adverbs, names of body parts, kinship terms and the number words from one to five; it avoids words specific to particular habitats or climates, as well as technical terms. We found that *dirty* was the most rapidly evolving word in the list, with a rate of lexical replacement of about 0.0009 per annum, or approximately one new non-cognate form every thousand years [1]. This rate of replacement yielded forty-seven different non-cognate forms among the 86 Indo-European languages in our sample. By comparison to words for *dirty*, the words with the slowest rates of lexical replacement were represented by just a single cognate form across the entire Indo-European language tree. Among these slowly evolving forms were the number words *two*, *three*, *five*, and the pronouns *who* and *I*.

The rates of lexical replacement for the slowly evolving words correspond to an expectation of one change in one hundred thousand years. If this figure seems extreme, consider that the Indo-European language family is somewhere between 7000 and 8000 years old [3]. Summing the time represented by all the branches that makes up the tree of the 86 Indo-European languages we studied yields approximately 140,000 language-years of potential evolution. Remarkably, during that time all the forms of the words for *two* (e.g., *dos*, *due*, *deux*, *duo*, and *twee*, among others) remained cognate, as did those for *three*, *five*, *who* and *I*. The words *one*, *four*, *we*, *when* and *tongue* round out the ten most slowly evolving words in the Indo-European languages.

The preponderance of number words in the list of slow evolvers raises the question of whether their slow rates of replacement are just an idiosyncrasy of Indo-European, or represent a more general phenomenon. Some reason to think that the slow rate of change of the number words might be a general phenomenon can be found in recent work on 'numerosity' – the ability to gauge number without a symbolic counting system – in animals. The ability to gauge number is almost certainly

1
2
3 useful in foraging, competitive, navigation and mating situations, and studies of the brains of
4 animals ranging from insects [4] to cephalopods [5], fish [6], amphibians [7], birds [8] and mammals
5 [9, 10] suggest the existence of dedicated populations of neurons attuned to the perception of
6 number, especially small numbers.
7

8 Here we extend our study of rates of lexical replacement from our previous Indo-European sample
9 to new data on the Indo-Europeans, and to Bantu and Austronesian language data sets, with special
10 emphasis on the relative rates of replacement of the 'low limit' number words *one* to *five*.
11

12 **2. Materials and Methods**

13 **(a) Lexical datasets**

14 We use three published lexical datasets. The Indo-European (IE) data comprise the words for 200
15 meanings in each of 103 languages [3]. The Austronesian data comprise 210 meanings and 400
16 languages [11], and here we use the 154 meanings with fewer than 200 cognate classes (see
17 Supplementary Materials). The Bantu data comprise 424 languages and 102 meanings [12]. The
18 meanings in these datasets are taken principally from the Swadesh fundamental vocabulary 200-
19 word list [2]. The raw data for the IE and Austronesian languages are available upon request from
20 the authors of those studies, and for the Bantu they are made available as part of the supplementary
21 information to that paper. Alternatively the IE data are available at IELex (ielex.mpi.nl) and the
22 Austronesian data are made available in the Austronesian Basic Vocabulary Database (ABVD,
23 language.psy.auckland.ns/austronesian).
24
25
26
27

28 **(b) Phylogenetic trees**

29 We used the Bayesian posterior samples of phylogenetic trees made available upon request by the
30 authors of the Indo-European and Austronesian, and for the Bantu as part of the supplementary
31 material of the original study [3, 11, 12]. Each study employed Bayesian Markov Chain Monte Carlo
32 methods [13] to estimate posterior distributions of time-calibrated trees. The trees are rooted and
33 have node ages derived from historical calibration points and statistical inference: the Indo-
34 European tree is dated to approximately 7654 +/- 915 years old, the Austronesian tree to 6924 +/-
35 500 years, and the Bantu tree to 6929 +/- 418 years. Branch lengths on the trees are calibrated in
36 years and so lexical replacement rates we report here are in units of expected changes per annum.
37
38

39 **(c) Cognate classifications**

40 The lexical datasets group the words for each meaning into between 1 and k cognate classes
41 denoting sets of words that are derived from a common ancestral word, based on expert linguist
42 judgments as described in the original references.
43
44

45 **(d) Modelling rates of lexical replacement**

46 Given the lexical data for each meaning coded into k distinct cognate classes, we observe for each
47 meaning a set of states ($1\dots k$) at the tips of phylogenetic tree T , where the tips correspond to
48 individual languages and the tree describes the patterns of descent of the set of languages from a
49 common ancestor (e.g., Figure 1).
50

51 We wish to discover the rates at which those k states arose given the assumption that they began
52 from a common ancestral state at the root of the tree. We presume that a series of replacements
53 has taken place throughout the tree eventually producing the k cognate sets. To capture this process
54 we define the instantaneous transition rates q_{jk} from any beginning state (cognate class) j to any end
55 state k , for all pairs of beginning and end states jk .
56

57 The set of q_{jk} defines a square matrix \mathbf{Q} of order $k \times k$, where \mathbf{Q} is given by
58
59
60

$$\mathbf{Q} = \begin{bmatrix} \dots & q_{12} & q_{13} & \dots & q_{1k} \\ q_{21} & \dots & q_{23} & \dots & q_{2k} \\ q_{31} & q_{32} & \dots & \dots & q_{3k} \\ \dots & \dots & \dots & \dots & \dots \\ q_{j1} & q_{j2} & q_{j3} & \dots & \dots \end{bmatrix},$$

and, by convention, the main diagonal elements (q_{jj}) are given by $-\sum q_{jk}$.

We expect k to vary considerably across meanings (e.g., compare $k=47$ for *dirty* and $k=1$ for *two* in our previous study), leading to the expectation of different average rates of lexical replacement among meanings. Accordingly, we re-write \mathbf{Q} as

$$\mathbf{Q} = r_i \frac{1}{c} \begin{bmatrix} \dots & q_{12} & q_{13} & \dots & q_{1k} \\ q_{21} & \dots & q_{23} & \dots & q_{2k} \\ q_{31} & q_{32} & \dots & \dots & q_{3k} \\ \dots & \dots & \dots & \dots & \dots \\ q_{j1} & q_{j2} & q_{j3} & \dots & \dots \end{bmatrix},$$

where r_i is now meaning i 's generalized rate of transition and the term $1/c$ is a normalization constant that, without any loss of generality, scales the q_{jk} to have a mean rate of 1.0. This scaling means that the q_{jk} can be interpreted as deviations around the generalized rate r_i . The normalization constant is calculated as

$$1/\sum_{jk} p_j q_{jk},$$

where p_j is the probability of state j in the observed data.

With \mathbf{Q} defined this way, the probability of a lexical change (appearance of new non-cognate word) from state j to state k over short interval of time dt , $\mathbf{P}_{jk}(dt) = \mathbf{Q}_{jk}dt$, where \mathbf{P} and \mathbf{Q} are matrices. To estimate the probability of transitions over longer time t , $\mathbf{P}_{jk}(t)$, \mathbf{Q} is exponentiated to give $\mathbf{P}(t) = e^{\mathbf{Q}t}$, and \mathbf{P} is the matrix of transition probabilities. This structure defines the usual continuous time Markov Model (e.g., [14]).

We estimate \mathbf{Q} using Bayesian Markov Chain Monte Carlo methods (e.g., [13]) to find

$$L(D/Q, T) = \int_{\mathbf{Q}, T} P(D/Q, T) d\mathbf{Q} dT,$$

where $L(D/Q, T)$ is the likelihood of the data (the observed cognate sets for a meaning) given \mathbf{Q} and the phylogenetic tree T . The Monte Carlo integration is performed simultaneously over increments in \mathbf{Q} and T and these increments are drawn from, in the case of \mathbf{Q} , a suitable proposal mechanism for altering the values of the q_{jk} , and in the case of T by calculating the likelihood over the posterior sample of trees. Integrating over \mathbf{Q} and T ensures that the estimates of the q_{jk} take into account uncertainty in the model of evolution and in the phylogenetic tree.

1
2
3
4 The number of elements in \mathbf{Q} increases as the square of k , the number of cognate sets. Thus, even
5 for a relatively small k , there can be a large number of parameters to estimate. To reduce the
6 severity of this problem we employ a reversible-jump Markov Chain Monte Carlo method we have
7 previously developed [15] that automatically collapses the large number of parameters in \mathbf{Q} into a
8 smaller number of distinct classes within which the individual q_{jk} can be regarded as identical
9 statistically.

10
11 The procedures for estimating the likelihood are implemented in the *BayesTraits* comparative-
12 phylogenetic analysis package (www.evolution.reading.ac.uk). We provide a sample command file in
13 the Supplementary Materials. The analysis yields a Bayesian posterior sample of \mathbf{Q} and the r_i , as
14 defined above. Our interest here is in the mean of the posterior sample of the r_i as an estimate of
15 the generalized rate of change for meaning i .
16

17 (e) estimation of a lexical half-life

18 Given a generalized rate r_i for meaning i , define the half-life of words for that meaning as the
19 expected amount of time before there is a 50% chance that word j will have been replaced by word
20 k [1, 16]. The half-life can be written as
21

$$22 t_{50} = -\text{Log}_e(0.5)/r_i$$

23 3. Results

24 (a) rates of lexical replacement in the three language families

25
26 The distribution of generalised rates over the Swadesh list items takes a broadly similar uni-modal
27 form in all three language families (Figures 2a-c), and rates of lexical replacement vary within each
28 family from 10 to over 100-fold (Table 1). The rate of replacement for *bird* in the Indo-European
29 languages at 0.00017 (Figure 1) falls just below the mean IE rate, and, as before, *dirty* has the fastest
30 rate of replacement. The Indo-European rates of change correlate $r=0.91$ with the rates of change
31 from our previous study [1] despite the new rates coming from a new tree that includes about 15%
32 more languages. Rates of change correlate strongly, but not perfectly, with the number of cognate
33 sets (a large number of cognate sets implies more replacements per unit time): $r=0.89$ for Indo-
34 European; $r=0.86$ for Bantu; $r=0.85$ for Austronesian (Figure 3a-c).
35

36 The lack of a perfect correlation between the number of cognate sets and rates of change illustrates
37 the importance of the phylogenetic tree, or more generally, of history in understanding evolution.
38 Two meanings might have an equal number of cognate sets but if the historical lexical replacements
39 (e.g., Figure 1) are distributed differently throughout the tree, the rates of replacement will also
40 differ. This is why the scatter about the regression lines in Figures 3a-c increases with the number of
41 cognate sets – as the number of cognate sets grows there are more different ways to distribute
42 them around the tree. For the data we report here, the phylogeny is responsible for around 21% to
43 28% of the variation in rates of change among the meanings, these figures being derived from $1-r^2$ of
44 the above correlations.
45

46 The average rates of lexical replacement in the IE and Bantu languages correspond to roughly a 20%
47 probability of lexical replacement per thousand years, remarkably close to the value Morris Swadesh
48 proposed in the 1950s from analysing differences between pairs of ancestral and descendant
49 languages – such as ancient and modern Greek -- separated by known times [2].
50
51
52
53
54
55
56
57
58
59
60

1
2
3 The average rate of lexical replacement among the Austronesian meanings is significantly higher
4 than for Bantu or IE. We cannot be certain whether this represents a true difference or perhaps a
5 difference in linguistic practice in identifying cognate words, the so-called 'lumpers' versus 'splitters'
6 problem that can also plague taxonomic practice in zoology. Alternatively, the Austronesian
7 expansion into Oceania was a process of 'island hopping' as the Austronesian people pushed further
8 and further into the unknown and uncharted Pacific [17]. It is possible then that serial founder
9 effects have influenced the Austronesian languages[18], where idiosyncrasies among the speakers
10 on a temporally ancestral island get magnified among the small number of speakers who move on to
11 descendant islands. Whatever the explanation, by restricting ourselves to the 154 meanings with
12 fewer than 200 cognate classes (see Materials and Methods) our average rate of lexical replacement
13 for Austronesian could even be an underestimate.
14

15
16 It is difficult to know why the upper bound of the Bantu rates is lower than that for IE or
17 Austronesian. It might reflect sampling: the 102 meanings in the Bantu list do not include the nine
18 fastest rate items from the IE list. On the other hand, rates of lexical replacement, while significantly
19 correlated among language families, are only modestly so: For IE and Austronesian, $r = 0.47$,
20 $p < 0.0001$; Bantu and Austronesian, $r = 0.37$, $p = 0.0006$; IE and Bantu, $r = 0.24$, $p = 0.0283$.
21

22 Half-life figures based on the rates of lexical replacement vary widely but even among the
23 Austronesian languages a slowly evolving word has a half-life of over 10,000 years (Table 1). The IE
24 languages seems to be extreme and this might arise because their smaller sample size and total tree
25 length mean some changes have been missed. The total tree length is the sum of the times over all
26 of the branches of the phylogenetic tree. For IE this is 148,400 years, for Bantu it is 490,660 and for
27 Austronesian it is 718,000.
28

29 30 (b) rates of lexical replacement of the low-limit number words (*one to five*) 31

32 The low-limit number words fall at the slower (lower) end of all three distributions of rates (Figure
33 2a-c, Table 1), and dominate the list of slowly evolving words in all three language families (Table 2).
34 Their rates of replacement are 3.5 to 20 times slower than the average rates of replacement and 10
35 to 130 times slower than the fastest rates of replacement (Table 1). Accordingly, low-limit number
36 words account for most of the longest half-lives (Table 1). Replacement rates for the number word
37 *one* are higher in all three languages families than for *two to five* (Table 2). We do not know why this
38 is the case but speculate that it might have something to do with *one* being replaceable in some
39 circumstances by 'a' or 'an'. This grammaticalisation by *one* to take over the use of articles has
40 occurred, among other languages, in English, German, Romanian, Spanish, French and Italian. The
41 probabilities of observing all five low-limit number words among the slowest eleven words, or four
42 of the five as in the case of Bantu and Austronesian, are all less than 0.0004 (Table 2). The extreme
43 slow rates of lexical replacement in the IE languages for the low-limit number words might arise
44 because 148,400 language-years is not sufficient to observe more than one change (as above).
45
46

47 48 c) low-limit number words in the Pama-Nyungan family 49

50 The Pama-Nyungan language family is widely geographically distributed throughout Australia [19,
51 20]. Its languages typically have simple low-limit number systems often not exceeding five [19]. A
52 dated phylogenetic tree for this language family is not available, making it impossible to calculate
53 lexical replacement rates. However, Claire Bowern (personal communication) who has studied this
54 group extensively has made available to us data for 183 vocabulary words in 190 Pama-Nyungan
55 languages, recording the rank orders across meanings of the number of cognate sets per meaning,
56 and classified into seventeen categories, including the number words, kinship terms and words for
57
58
59
60

1
2
3 the environment. The dataset includes three number words – *one*, *two* and *three* – and their mean
4 rank order is the lowest (fewest cognate sets) of any of the seventeen categories of words.
5

6 7 **4. Three hypotheses to explain the unusual conservation of the number** 8 **words** 9

10 Previously we have shown that words used more frequently in everyday discourse tend to be among
11 the most conserved or slowly evolving [1]. Even among the slowly evolving words, the number
12 words are unusual in having rates of lexical replacement considerably slower than would be
13 predicted from their frequency of use [1]. Here we speculate on three hypotheses that might
14 explain why the number words evolve so slowly, and offer data consistent with each.
15

16 (a) evolutionarily conserved brain regions associated with numerosity (somehow) influence
17 the learning and use of linguistic-symbolic number words
18

19
20 Could the evolutionarily ancient and seemingly hard-wired nature of many animals' abilities to
21 perceive 'number' independently of a symbolic language for counting ([4], [5],[6],[7], [8],[9],[10]) be
22 linked to the slow replacement rate of number words? Brain regions associated with numerosity are
23 distinct from those involved in language [10, 21]. Still, brains are vast interconnected and highly
24 parallel networks that can make available their internal representations or outputs to other brain
25 regions. Perhaps an unambiguous brain state associated with simple judgements of different
26 numbers of objects – so called numerosity judgements – makes number words easier for humans to
27 learn or strengthens the association of numerosity to the symbolic number words, thereby slowing
28 their rates of replacement.
29

30
31 Data from a study of the age of acquisition for 30,000 English words [22] might be relevant to this
32 idea. Children learn words earlier the more frequently those words are used in common everyday
33 speech. But using the Kuperman et al. [22] data, we find that all ten number words from *one* to *ten*
34 have earlier ages of acquisition than is predicted from their frequency of use (binomial test, $p < 0.002$,
35 two-tailed; Figure 4).
36

37 (b) number words are unambiguous in their meanings and therefore less likely to admit
38 alternatives
39

40
41 If the number words are unambiguous in their meanings, or at least relatively so compared to other
42 meanings, then speakers might be less likely to use alternatives for them in everyday speech. For
43 example, shown three objects and asked to describe 'how many', speakers will overwhelmingly say
44 'three'. But speakers describing, for example, a weather storm that includes thunder and lightning
45 might call it a *thunderstorm* or *thunder and lightning* or perhaps a *lightning storm*. Each of these
46 alternative forms is likely to be understood and thereby might be allowed to co-exist in the
47 population of speakers.
48

49
50 If it is generally true that the number words admit fewer alternatives, then, from a population-
51 genetic perspective the mutation rate (rate at which new words enter the lexicon) for number words
52 is lower than the mutation rate for other kinds of words. The neutral theory of evolution [23]
53 demonstrates that the rate of evolution of neutral alleles is equal to the rate of neutral mutation. If
54 we entertain the possibility that alternative words for a meaning might be equally good – and
55 therefore neutral -- then the lower mutation rate of number words predicts their slower rate of
56 lexical replacement.
57
58
59
60

1
2
3 Large-scale surveys that record the words people use in conversation [24] reveal that for some
4 common objects and actions a variety of different words might be used, whereas for others most
5 respondents use the same word: days of the week, months of the year and the number words fall
6 into this latter category.

7 Brysbaert et al. [25] provide ratings of ‘concreteness’ for 40,000 English words. The number words
8 for *one* to *ten* receive a mean concreteness rating (five-point scale) of 3.78 ± 0.33 (s.e.m., $n=10$),
9 significantly higher than the overall mean of 3.04 ± 0.005 ($n=39,894$), although not significantly higher
10 than nouns (3.53 ± 0.008 , $n=14,592$). But the Brysbaert ‘concreteness’ scale measures “things or
11 actions in reality, which you can experience directly through one of the five senses”, and so is not
12 directly relevant to the sense we are suggesting here of ‘unambiguous’, corresponding to a meaning
13 for which, owing to the unambiguous nature of the concept, only a single word generally applies.
14 Thus, the highest scoring words in the Brysbaert sample included ‘spaghetti sauce’, ‘trench coat’,
15 ‘thorn’ and ‘angelfish’, all of which received a score of five but for which one can easily imagine
16 alternative words.
17

18
19 (c) number words occupy a region of the phonetic space that is relatively full
20

21 Shorter words define a smaller space of possible words than longer ones. The exact size of the space
22 of possible words will depend upon a language’s phonotactic rules [26] governing permissible
23 combinations of sounds. For instance, no English word begins with the velar nasal sound *ng*,
24 although this combination is common in other languages and occurs at the end of many English
25 words. If the phonotactic rules could be known precisely for a language it would be possible to
26 generate all of the possible words of a given length for that language. But even without knowing
27 what these rules are, the space of possible words will grow rapidly, probably something close to
28 factorially, with a word’s length.
29

30
31 Data from the British National Corpus [27] record the frequency of use of thousands of common
32 words in everyday speech and writing. These data reveal that a word’s length (scored conservatively
33 here as the number of letters rather than the number of distinct sounds) declines sharply with its
34 frequency of use (Figure 5). Zipf [28] had already been identified this relationship by the late 1940s
35 when he put forward his principle of least effort to explain, among other things, why the frequently
36 used words became shorter.
37

38
39 If we accept Zipf’s principle, then words will continually evolve to become shorter, and the more so
40 the more they are used. It might just be, then, that the pressure to become shorter means that the
41 already smaller phonetic space of shorter words is full or nearly full compared to the space for
42 longer words. If the space is full, then possible replacements for a word already in that small space
43 might in general have to be longer, or more difficult to pronounce and in that sense not as ‘fit’ as the
44 original. This lower fitness might make the word less likely to be adopted, and as a consequence
45 would slow the rate of lexical replacement.
46

47
48 Anecdotally, the phonetic space for short words can seem full. Compare the words *two*, *to*, *too* and
49 *you* in English or *deux*, *tu* and *vous* in French. These words, all highly used, have crowded in on each
50 other, occupying nearly identical phonetic spaces. In the extreme this crowding produces
51 homophones, words with the same sound but different meanings, such as *pale* and *pail*. An
52 analogous concept in genetics is alternative splicing [29] whereby a single gene can produce more
53 than one protein. Alternative splicing allows an organism to produce many more different proteins
54 than would be expected from its number of genes, and can be seen as a way organisms can reduce
55 the amount of DNA they have to carry and reproduce.
56
57
58
59
60

1
2
3 A prediction consistent with the phonetic-space-full argument is that homophones should, in
4 general, be shorter words than non-homophones, reflecting the pressure for words to become
5 shorter but having a smaller phonetic space to occupy. To test this idea we recorded the average
6 length of 441 pairs and triples of British English homophones and then compared the average length
7 of these homophones to the length of words in the British National Corpus (Figure 6). The
8 homophones are significantly shorter: mean homophone length = 4.56 ± 0.041 (s.e.m.), $n=441$ pairs
9 and triples or 991 words total; mean length BNC = 6.93 ± 0.029 , $n=6956$ (7.08 ± 0.03 excluding
10 homophones), $p < 0.0001$.
11

12
13 There can be disagreement about whether two or more words are homophones (e.g., *all* and *awl* or
14 *close* and *clothes*) and it might be more difficult to form homophones of longer words (although the
15 many more possible long words might offset this), but the result in Figure 6 is consistent with the
16 idea that the vast space of possible long words makes homophones of them less necessary because
17 there are so many more possible alternatives from which to choose. Nevertheless, a challenge for
18 the ‘phonetic-space-is-full’ argument as an explanation for the number words is that it applies
19 equally to all short words, including the pronouns and the “wh” words (*who*, *what* *where*, *why* and
20 *when*). These words are also among some of the most slowly evolving in the Indo-European
21 languages ([1]; Table 2 this paper) and frequency data show that they are all highly used. But among
22 these slowly evolving words, the rate of lexical replacement for the number words is exceptionally
23 slow even for their frequency of use [1]. This does not necessarily invalidate the phonetic space
24 argument, but signals that there might be some additional factor slowing replacement rates of the
25 number words.
26

27 28 **5. Discussion**

29
30 There does seem to be something special about the number words: at least in the three language
31 families we studied, the low-limit number words have unusually slow rates of lexical replacement,
32 meaning that a shared form of the word can often last many thousands of years. The same also
33 seems true of the Pama-Nyungan language family of Australia. We speculated upon three reasons
34 why the number words have low rates of lexical replacement, and offered some evidence consistent
35 with each. More work on each of these hypotheses would be a welcome addition to understanding
36 the beguiling stability of the number words.
37

38
39 In contrast to the unusual conservation of the low-limit number words (and especially *two* to *five*)
40 higher level number words such as the ‘teens’ (in English 13 to 19) and the names of the numbers
41 that are powers of ten can be more variable ([30]). The form these higher-level number words take –
42 for example, sometimes adding a base number to ten, sometimes adding *ten* to a base number –
43 correlates with features of a language’s grammar [30]. This greater variation and the association
44 with grammar may indicate that the higher-level number words are relatively recent inventions, or
45 put another way, that the low-limit number words are culturally ancestral, existing from a time
46 when counting above small numbers was unusual or unnecessary. Indeed, some hunter-gatherer
47 languages are claimed to lack number words altogether [31] [32] [33] [34]. Alternatively, the
48 combinatoric nature of the higher number words might make them inherently more prone to
49 change.
50

51
52 Some words we might expect to be highly conserved are not. Names of body parts, and relational
53 words for *mother*, *father*, *husband* and *wife*, or *he* and *she*, or perhaps words for *fire* or *spears* might
54 all be expected to play central roles in everyday speech and especially so in ancient societies, and
55 therefore be conserved. But with the exception of *child*, *eye* and *tongue* none of these words made
56 it into the slowest evolving set of words (Table 2) for any of the language families. Indeed, in
57 contrast to the extreme conservation of the number words, there are forty-three different cognate
58
59
60

forms of the words for *husband* in the Indo-European languages, and thirty-seven of the words for *wife*.

It is worth putting into a temporal context the extraordinary conservation of some of the number words. In the Indo-European languages, the numbers words for *two*, *three* and *five* are all represented by a single cognate set. The Indo-European language tree we used has a total tree length spanning 148,400 language-years. For a word to remain cognate among the languages of the Indo-European tree means that every speaker of its many languages used a cognate form of that word throughout history, or at least if some other forms were tried, they never caught on. Words that can live this long should astound us, because there were no writing systems for nearly all of the history of the Indo-European language family and the opportunities are great for an aural signal to be corrupted: when a speaker utters a word that sound travels as a pressure wave through the air where it is transduced by a listener's ear into an electrical signal that travels to the brain and is stored in some memory state. Then, when that speaker uses the same word it must be transformed back from the stored brain state into a set of instructions to the facial muscles, lungs and abdomen of the speaker to form the pressure wave anew. That this process can be repeated millions or perhaps billions of times throughout history with so little change cries out for an understanding of how our minds achieve this prodigious feat.

Acknowledgments

We thank Annemarie Verkerk for discussion of the phonetic space argument and Andreea Calude for examples of the possible grammaticalisation of the word 'one'.

Data Accessibility

The rates datasets we derived for the analyses in Figures 2 and 3 and Tables 1 and 2 are available as Supplementary Material to this article. The raw lexical data and the posterior samples of phylogenetic trees are available from the authors of the original articles on the IE, Austronesian and Bantu languages. The age of acquisition data are available from the authors of that study. The British National Corpus data are available online.

Authors' Contributions

MP and AM both contributed to all aspects of the paper.

Competing Interests

We have no competing interests.

Funding

This work was supported by Advanced Investigator Award 268744 'MotherTongue' to M.P. from the European Research Council, and by BBSRC grant number BBSRC grant BB/L018594/1 to A.M.

References

1. Pagel M., Atkinson Q.D., Meade A. 2007 Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* **449**(7163), 717-720.
2. Swadesh M. 1952 Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proceedings of the American philosophical society* **96**(4), 452-463.
3. Bouckaert R., Lemey P., Dunn M., Greenhill S.J., Alekseyenko A.V., Drummond A.J., Gray R.D., Suchard M.A., Atkinson Q.D. 2012 Mapping the origins and expansion of the Indo-European language family. *Science* **337**(6097), 957-960.
4. Dacke M., Srinivasan M.V. 2008 Evidence for counting in insects. *Animal cognition* **11**(4), 683-689.

- 1
 - 2
 - 3
 - 4
 - 5
 - 6
 - 7
 - 8
 - 9
 - 10
 - 11
 - 12
 - 13
 - 14
 - 15
 - 16
 - 17
 - 18
 - 19
 - 20
 - 21
 - 22
 - 23
 - 24
 - 25
 - 26
 - 27
 - 28
 - 29
 - 30
 - 31
 - 32
 - 33
 - 34
 - 35
 - 36
 - 37
 - 38
 - 39
 - 40
 - 41
 - 42
 - 43
 - 44
 - 45
 - 46
 - 47
 - 48
 - 49
 - 50
 - 51
 - 52
 - 53
 - 54
 - 55
 - 56
 - 57
 - 58
 - 59
 - 60
5. Yang T.-I., Chiao C.-C. 2016 Number sense and state-dependent valuation in cuttlefish. In *Proc R Soc B* (p. 20161379, The Royal Society).
6. Agrillo C., Dadda M., Serena G., Bisazza A. 2008 Do fish count? Spontaneous discrimination of quantity in female mosquitofish. *Animal cognition* **11**(3), 495-503.
7. Rose G.J., Leary C.J., Edwards C.J. 2011 Interval-counting neurons in the anuran auditory midbrain: factors underlying diversity of interval tuning. *Journal of Comparative Physiology A* **197**(1), 97-108.
8. Rugani R., Cavazzana A., Vallortigara G., Regolin L. 2013 One, two, three, four, or is there something more? Numerical discrimination in day-old domestic chicks. *Animal cognition* **16**(4), 557-564.
9. Nieder A., Freedman D.J., Miller E.K. 2002 Representation of the quantity of visual items in the primate prefrontal cortex. *Science* **297**(5587), 1708-1711.
10. Harvey B., Klein B., Petridou N., Dumoulin S. 2013 Topographic representation of numerosity in the human parietal cortex. *Science* **341**(6150), 1123-1126.
11. Gray R.D., Drummond A.J., Greenhill S.J. 2009 Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *science* **323**(5913), 479-483.
12. Grollemund R., Branford S., Bostoen K., Meade A., Venditti C., Pagel M. 2015 Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proceedings of the National Academy of Sciences* **112**(43), 13296-13301.
13. Gilks W.R., Richardson S., Spiegelhalter D.J. 1996 Introducing markov chain monte carlo. *Markov chain Monte Carlo in practice* **1**, 19.
14. Pagel M. 1994 Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London B: Biological Sciences* **255**(1342), 37-45.
15. Pagel M., Meade A. 2006 Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *The American Naturalist* **167**(6), 808-825.
16. Pagel M. 2000 The history, rate and pattern of world linguistic evolution. *The evolutionary emergence of language: social function and the origins of linguistic form*, 391-416.
17. Kirch P.V., Green R.C., Bellwood P.S., Dunnell R., Dye T., Gosden C., Rowe C.W., Terrell J., Vogt E.Z., Welsch R. 1987 History, Phylogeny, and Evolution in Polynesia [and Comments and Reply]. *Current Anthropology* **28**(4), 431-456.
18. Atkinson Q.D., Meade A., Venditti C., Greenhill S.J., Pagel M. 2008 Languages evolve in punctuational bursts. *Science* **319**(5863), 588-588.
19. Zhou K., Bower C. 2015 Quantifying uncertainty in the phylogenetics of Australian numeral systems. In *Proc R Soc B* (p. 20151278, The Royal Society).
20. Bower C., Atkinson Q. 2012 Computational phylogenetics and the internal structure of Pama-Nyungan. *Language* **88**(4), 817-845.
21. Dehaene S., Spelke E., Pinel P., Stanescu R., Tsivkin S. 1999 Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science* **284**(5416), 970-974.
22. Kuperman V., Stadthagen-Gonzalez H., Brysbaert M. 2012 Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods* **44**(4), 978-990.
23. Kimura M. 1984 *The neutral theory of molecular evolution*, Cambridge University Press.
24. Davis A.L. 1969 A Compilation of the Work Sheets of the Linguistic Atlas of the United States and Canada and Associated Projects.
25. Brysbaert M., Warriner A.B., Kuperman V. 2014 Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods* **46**(3), 904-911.
26. Hayes B. 2011 *Introductory phonology*, John Wiley & Sons.
27. Consortium B.N.C. 2007 British National Corpus version 3 (BNC XML edition). *Distributed by Oxford University Computing Services on behalf of the BNC Consortium Retrieved February 13, 2012.*
28. Zipf G.K. 1949 Human Behaviour and the Principle of Least-Effort. Cambridge MA edn. (Addison-Wesley, Reading).

- 1
2
3 29. Brett D., Pospisil H., Valcárcel J., Reich J., Bork P. 2002 Alternative splicing and genome
4 complexity. *Nature genetics* **30**(1), 29.
5 30. Calude A.S., Verkerk A. 2016 The typology and diachrony of higher numerals in Indo-
6 European: a phylogenetic comparative study. *Journal of Language Evolution* **1**(2), 91-108.
7 31. Gordon P. 2004 Numerical cognition without words: Evidence from Amazonia. *Science*
8 **306**(5695), 496-499.
9 32. Pica P., Lemer C., Izard V., Dehaene S. 2004 Exact and approximate arithmetic in an
10 Amazonian indigene group. *Science* **306**(5695), 499-503.
11 33. Frank M.C., Everett D.L., Fedorenko E., Gibson E. 2008 Number as a cognitive technology:
12 Evidence from Pirahã language and cognition. *Cognition* **108**(3), 819-824.
13 34. Bowern C., Zentz J. 2012 Numeral systems in Australian languages. *Anthropological*
14 *Linguistics* **54**, 130-166.
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For Review Only

Table 1. Average \pm standard deviations of lexical replacement rates and half-lives for fundamental vocabulary items and for low limit number words in the three language families.

replacement rate, per annum	Indo-European (n=200 words)	Bantu (n=102 words)	Austronesian (n=154 words)
overall	0.00020 \pm 0.00011	0.00023 \pm 0.00009	0.00035 \pm 0.00012
<i>fastest, slowest, ratio (f/s)</i>	0.00061, 0.0000047, 130	0.00045, 0.000026, 17	0.00065, 0.000065, 10
<i>half-life, years: average, shortest, longest,</i>	3465, 1066, 147,000	3150, 1540, 26,659	1980, 1066, 10,582
'low limit' number words (one to five)	0.00001 \pm 0.00004 (p<0.0001)	0.00011 \pm 0.00009 (p<0.003)	0.00016 \pm 0.00005 (p<0.0001)
exclude one		0.00006 \pm 0.00003 (p<0.00003)	0.00010 \pm 0.0.00005 (p<0.0001)

Table 2. Rank order of rate of lexical replacement for the eleven meanings with the slowest rates of change; rank = 1 is slowest. Words 'one' to 'five' in bold.

rank	Indo-European (n=200 words)	Bantu (n=102 words)	Austronesian (n=154 words)
1	two	eat	child
2	three	tooth	twoe
3	fiv	three	to pound/beat
4	who	eye	three
5	four	fiv	to die
6	l	hunger	eye
7	one	elephant	four
8	we	four	ten
9	when	persone	fiv
10	tongue	child	tongue
11	name	two	eight

Note: The probability of all five low limit number words appearing in the slowest eleven for IE is p= 0.0000002; the probability of four of the five low limit number words appearing in the slowest eleven for Bantu is 0.00036 and 0.00007 for Austronesian.

Figure and Table Captions

Figure 1. Partial phylogenetic tree of the Indo-European languages showing the words that the languages use for the meaning 'bird', coded to identify cognate classes. Squares along the branches identify regions of the tree where new cognate classes might have arisen, although the analysis strategy integrates over all possible ancestral transitions (Pagel, 1994) and so is not conditional upon any particular set of them.

Figure 2a-c. Rates of lexical replacement per annum. a) rates of lexical replacement in the Indo-European languages for 200 Swadesh list meanings; b) rates of lexical replacement in the Bantu languages for n=102 meanings; c) rates of lexical replacement in the Austronesian languages for n=154 meanings. The darker shaded areas of each histogram correspond to the position of the low limit number words (*one to five*). The rate for *one* is elevated in the Bantu and Austronesian datasets.

Figure 3a-c. Correlations between number of cognate sets (x-axis, NOS) and rate of lexical replacement, in a) the Indo-European languages for 200 Swadesh list meanings; b) the Bantu languages for n=102 meanings; c) the Austronesian languages for n=154 meanings. The darker circles correspond to the low limit number words. The rate and number of states for *one* are relatively high in the Bantu and Austronesian datasets.

Figure 4. Age of acquisition of a word versus frequency of use. The numbers words from 'one' to 'ten' (heavy black dots) all fall below (have earlier ages of acquisition) than expected from their frequency of use. Regression fit $r = 0.65$. Raw data taken from Kuperman et al. [22]

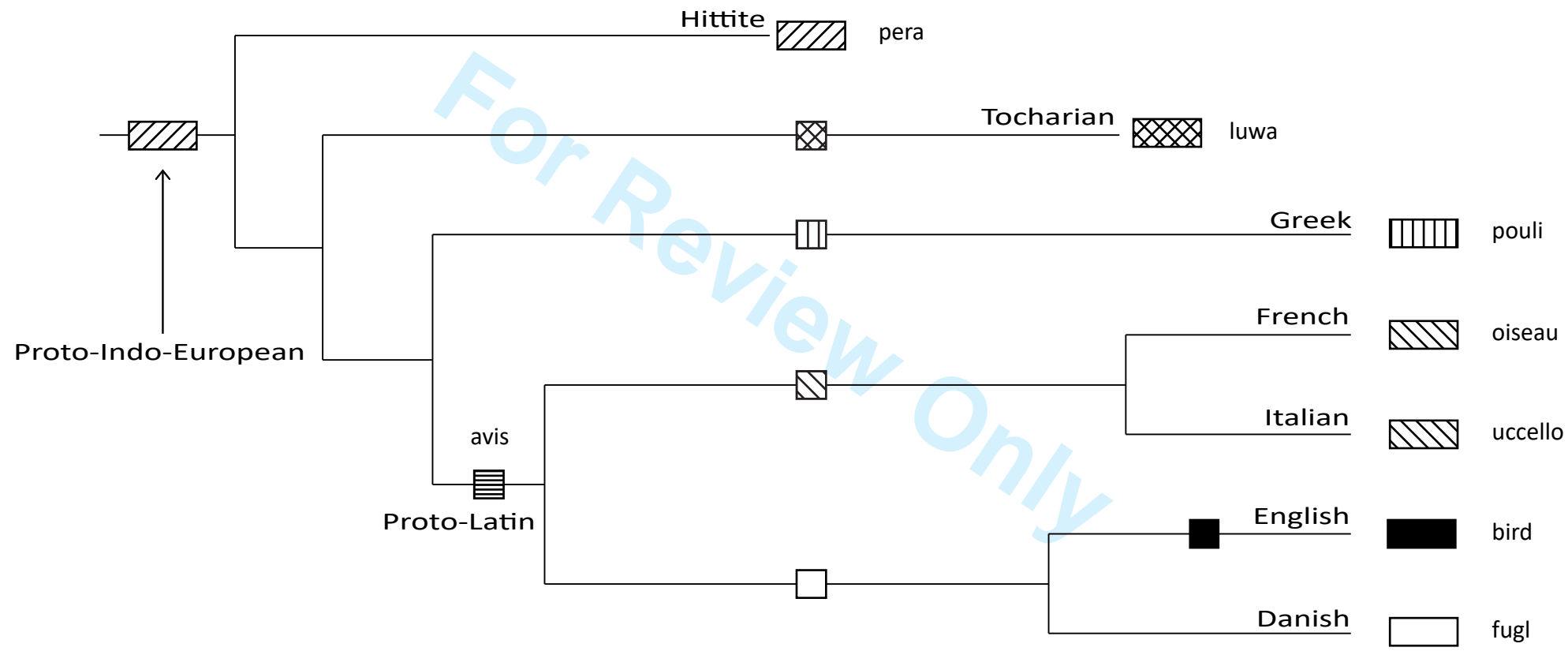
Figure 5. Length of word (in characters) versus frequency of occurrences per million from the ~7700 most frequently occurring words in the rank-ordered-by-frequency list of the British National Corpus (http://ucrel.lancs.ac.uk/bncfreq/lists/1_2_all_freq.txt). Note: x-axis truncated at 5000, frequencies extend to >60,000. Shaded area is region of the numbers words from 'one' to 'ten' (heavy black dots).

Figure 6. Frequency histogram of word length from the rank-ordered-by-frequency list of the British National Corpus (as in Fig. 4); shaded area is word length of homophones within the BNC sample. The BNC list includes all words down to a frequency of 10 per million, yielding n=7726 words. Removing abbreviations, proper nouns, names and special characters leaves n=6956 words. Mean length BNC = 6.93 ± 0.029 (s.e.m.), or 7.08 ± 0.03 excluding homophones; mean length of homophones = 4.56 ± 0.041 (s.e.m.), n=441 pairs and triples or 991 words total, $p < 0.0001$. Homophones taken from <http://www.singularis.ltd.uk/bifroest/misc/homophones-list.html>; a comparison sample of homophones is made available at <http://www.teachingtreasures.com.au/teaching-tools/Basic-worksheets/worksheets-english/upper/homophones-list.htm>: mean homophone length = 4.77 ± 0.045 , n=427 pairs.

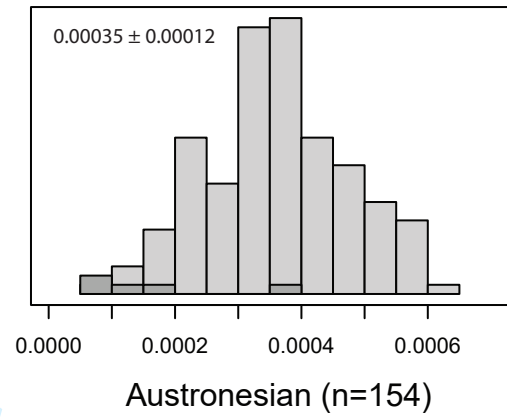
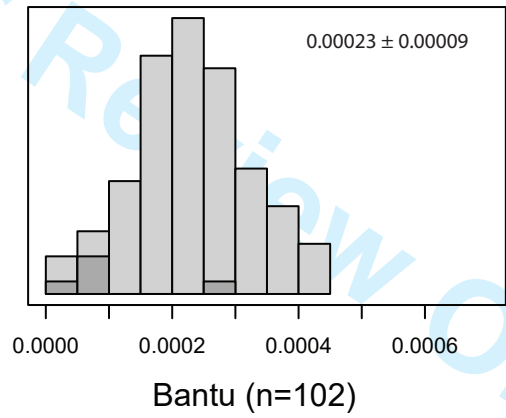
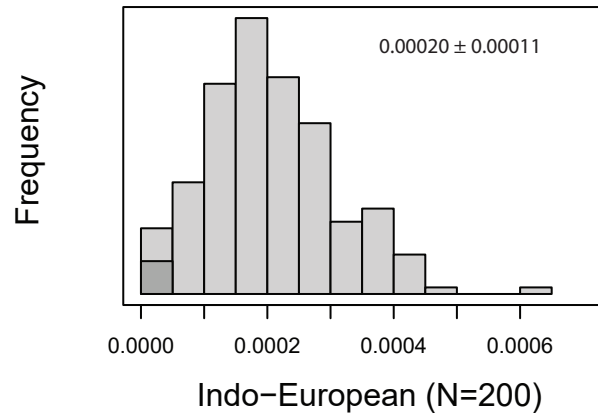
Table 1. Average lexical replacement rates and half-lives for Swadesh fundamental vocabulary meanings [2] and for low-limit number words in the three language families.

Table 2. Rank order of rate of lexical replacement for the eleven meanings with the slowest rates of change; rank = 1 is slowest. Words 'one' to 'five' in bold.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

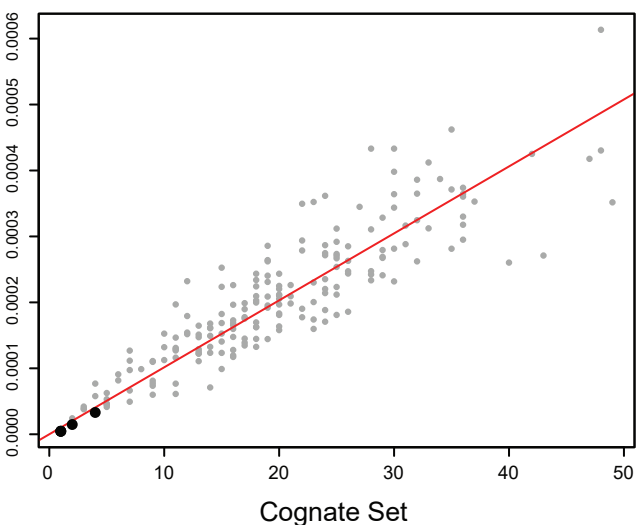


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

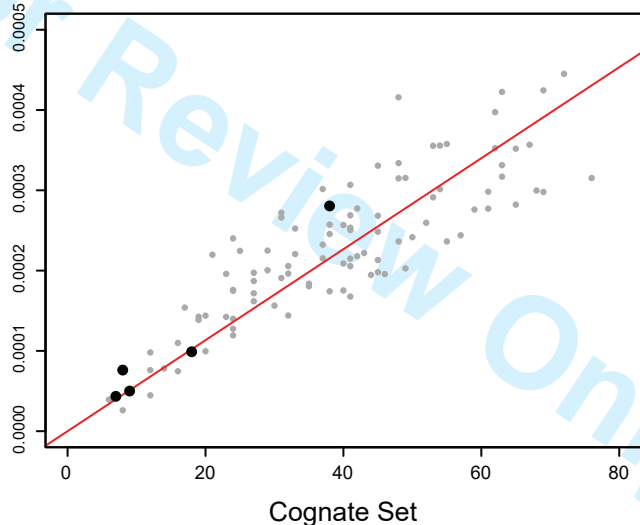


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

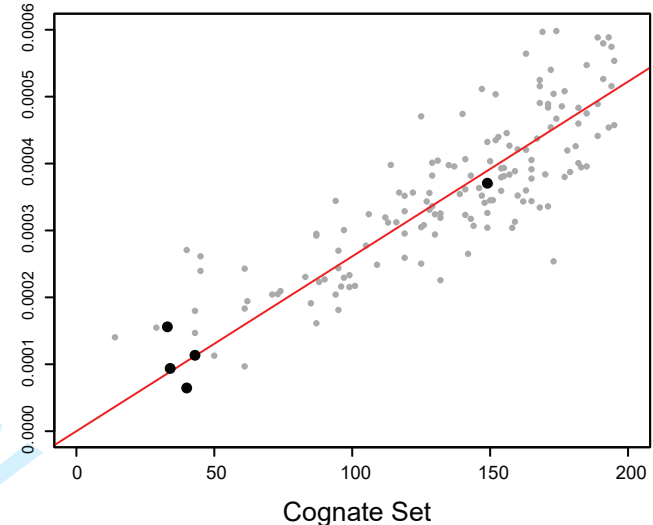
Indo-European

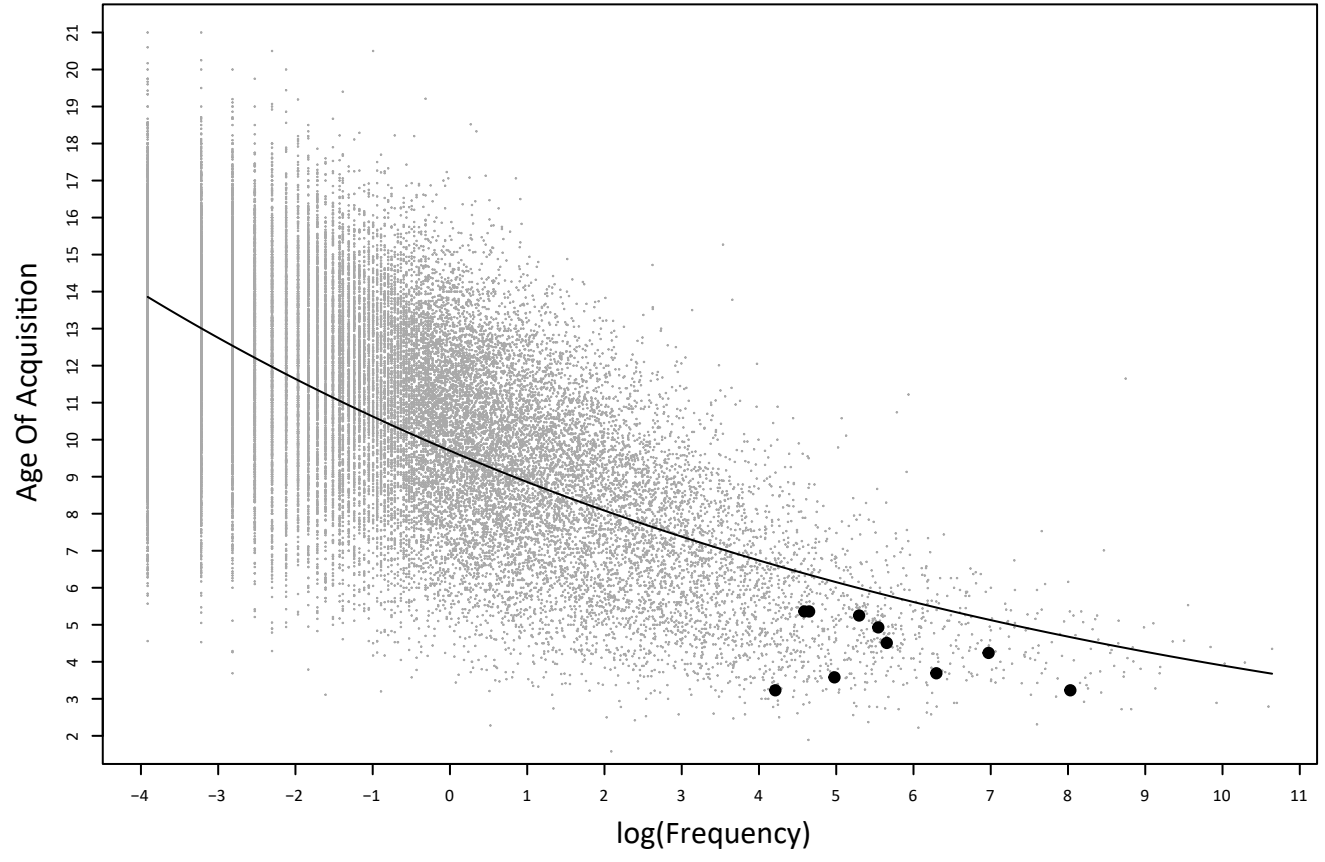


Bantu



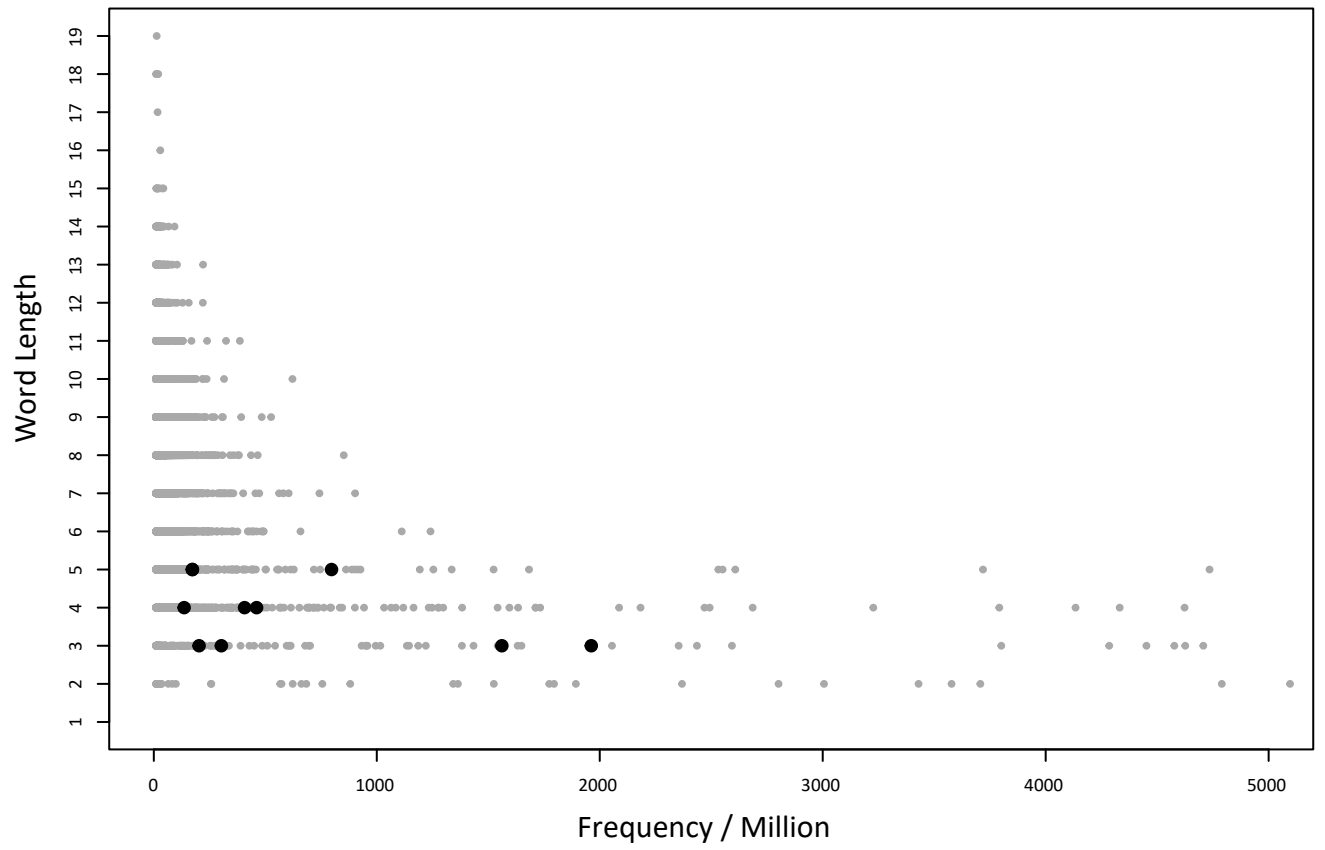
Austronesian

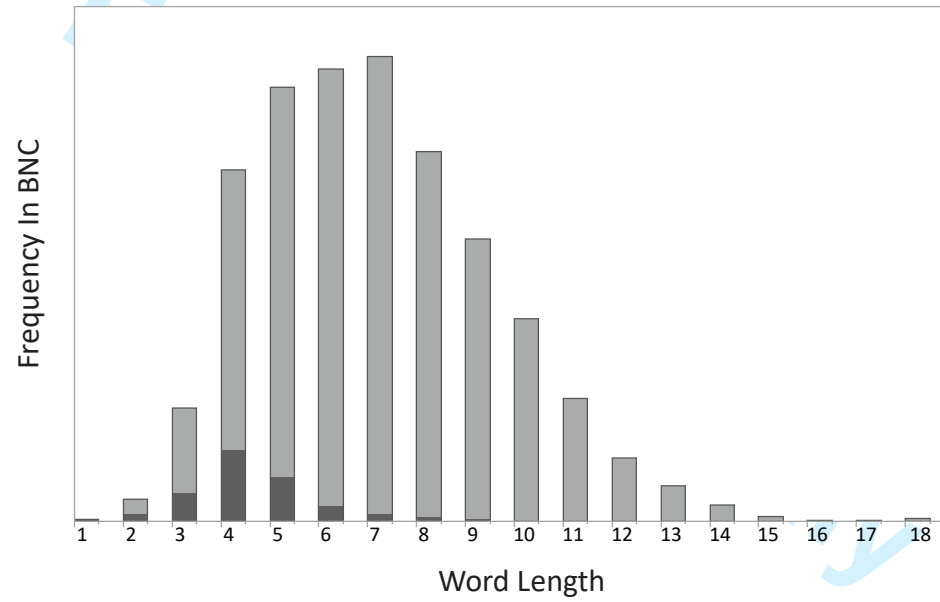




1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47





1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47