Dominant words rise to the top by positive frequency-dependent selection

Article

Accepted Version

Pagel, M. ORCID: https://orcid.org/0000-0001-7287-8865, Beaumont, M. A., Meade, A. ORCID: https://orcid.org/0000-0001-7095-7711, Verkerk, A. and Calude, A. (2019) Dominant words rise to the top by positive frequency-dependent selection. Proceedings of the National Academy of Sciences of the United States of America, 116 (15). pp. 7397-7402. ISSN 1091-6490 doi: https://doi.org/10.1073/pnas.1816994116 Available at https://centaur.reading.ac.uk/83133/

To link to this article DOI: http://dx.doi.org/10.1073/pnas.1816994116

Publisher: National Academy of Sciences

Dominant words rise to the top by positive frequency-dependent selection

Mark Pagel[1,2], Mark Beaumont[3], Andrew Meade[1], Annemarie Verkerk[1,†] and Andreea Calude[4]

1. School of Biological Sciences, University of Reading, Whiteknights, Reading RG6 6UR
2. Santa Fe Institute, 1399 Hyde Park Rd, Santa Fe, NM 87501, USA
3. School of Biological Sciences, Life Sciences Building, University of Bristol, 24 Tyndall Avenue, Bristol, BS8 1TW
4. Department of General and Applied Linguistics, University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand

† Current address: Max Planck Institute for the Science of Human History, Kahlaische Strasse 10, 07745 Jena, Germany

Classification: Biological Sciences, Evolution

Author for correspondence:
Mark Pagel
email: m.pagel@reading.ac.uk
phone: 0118 378 8900

Dominant words rise to the top by positive frequency-dependent selection

Abstract

A puzzle of language is how speakers come to use the same words for particular meanings, given that there are often many competing alternatives (e.g., sofa, couch, settee), and there is seldom a necessary connection between a word and its meaning. The well-known process of random drift – roughly corresponding in this context to 'say what you hear' – can cause the frequencies of alternative words to fluctuate over time, and it is even possible for one of the words to replace all others, without any form of selection being involved. But is drift alone an adequate explanation of a shared vocabulary? Darwin thought not. Here we apply models of neutral drift, directional selection and positive-frequency-dependent selection to explain over 417,000 word-use choices for 418 meanings in two natural populations of speakers. We find that neutral drift does not in general explain word-use. Instead, some form of selection governs word-choice in over 91% of meanings. In cases where one word dominates all others for a particular meaning – such as is typical of the words in the core lexicon of a language – word-choice is guided by positive-frequency-dependent selection – a bias that makes speakers disproportionately likely to use the words that most others use. This bias grants an increasing advantage to the common form as it becomes more popular and provides a mechanism to explain how a shared vocabulary can spontaneously self-organise, and then be maintained for centuries or even millennia, despite new words continually entering the lexicon.

Significance

Speakers of a language somehow come to use the same words to express particular meanings – like *dog* or *table* – even though there is seldom a necessary connection between a word and its meaning, and there are often many alternatives from which to choose (e.g., *sofa*, *couch*, *settee*). We show that word-choice is not just a matter of saying what others say. Rather, humans seem to be equipped with a bias that makes them disproportionately more likely to use the words that most others use. The force of this bias can drive competing words out, allowing a single word to dominate all others. It can also explain how languages spontaneously organise, and remain relatively stable for centuries or even millennia.

Introduction

In his review of August Schleicher's 1869 pamphlet *Darwinism Tested by the Science of Language* (1), the 19th century philologist Max Müller wrote "a struggle for life is constantly going on amongst the words and grammatical forms in each language. The better, the shorter, the easier forms are constantly gaining the upper hand, and they owe their success to their own inherent virtue" (2). Evidently, so taken was Darwin with Müller's views that just a year later he quoted Müller's "struggle for life…" passage in his 1871 book the *Descent of Man* (3), adding "the survival or preservation of certain favoured words in the struggle for existence is natural selection" (p91).

Linguists since Schleicher's time have continued to identify regularities in the ways that languages change, including patterns in the replacement of sounds, morphology, syntax and words (4-6). For instance, frequently used words tend to be replaced less often than infrequently used words (7) and irregular verbs have a greater tendency to become regular than do regular verbs to become irregular (8). Linguistic change such as in these two examples, involves some form of competition among alternative words, but were Müller and Darwin right to assume that the changes are driven by natural selection, that is to say, the changes are driven by the "inherent virtue" of the eventual winners?

One of the more significant developments of twentieth century neo-Darwinism was the mathematical formulation of the theory of neutral or random drift (9, 10). This theory, commonly applied to genetic variants, shows that the frequencies of alternative forms change over time simply as a result of random or stochastic effects – no selection need be involved. Applied to language (11, 12) random drift can be used to study changes in the frequencies with which speakers use various words for a given meaning, such as *sofa*, versus *couch* or *settee*. Drift's importance in population studies, then, is that its mathematical expression provides a precise null expectation against which stronger claims, such as those that Darwin and Müller made, can be assessed (11, 12).

For example, in language a common observation is that when the number of speakers who use a word is plotted against that word's rank order position in a list of words sorted by frequency, sharply down-sloping curves arise that can be described by the form $f(k) = \alpha k^{-\beta}$, where $f(k)$ is the observed number of speakers who use a word, and k is its rank order position (*1,2,….k*) (13). Figures 1a-c plot this relationship for the number of people who use one of the k alternative words for a given meaning. Studies in linguistic settings have shown that drift can produce curves with these shapes (12, 14-16), even the extreme example in Figure 1c where, among competing alternatives, one word has risen to the top dominating all others. On the other hand, while drift can in principal produce any monotonically declining curve, some outcomes of drift are more probable than others (17), and in particular shapes such as Figures 1b and c are relatively unlikely under drift. So, the real question becomes not whether drift can produce outcomes such as those in Figures 1a-c, but whether mechanisms other than drift provide more likely explanations. This is the challenge that claims of selection in language must meet.

Here we investigate the contributions of random drift (D) along with three forms of selection – directional selection (DS), positive frequency-dependent selection (FDS) and a model that combines directional with positive frequency-dependent selection (DS+FDS; **Methods**). Drift asks what frequency distributions of speakers per word emerge over long periods of time if speakers use words randomly in proportion to the number of other speakers using them. Directional selection incorporates drift but allows some words to be inherently better or worse than others. An example of directional selection is that shorter words, or words that are easier to pronounce might have an advantage, especially when they are frequently used in speech (18). Alternatively, a word might acquire an advantage from being used by a high status person. Directional models including various social, phonetic or other biases have been proposed for linguistic change (19), or, for example, in cultural settings to understand the choice of colour terms, musical preferences or baby names (20).

Positive frequency dependent selection refers to a scenario in which the likelihood that a speaker will use a word increases disproportionately to the number of other speakers using it. Positive frequency dependence is observed in nature for aposematic or warning colours in insects, as the aposematic signal often becomes increasingly effective at deterring predators as it spreads through a population (21). Elements of the frequency-dependent process appear in early work in statistics (22), and in cultural settings, positive frequency-dependence or 'conformist bias' (23) has been investigated to explain the evolution of cultural forms (24), and the diffusion of innovations (25).

We implement these models in a computational framework that allows us to assess their relative contributions to explaining word choice in two regional American populations.

Data and Results

Our data come from over 417,000 responses obtained from over 2000 respondents in two regional surveys conducted as part of the Linguistic Atlas Project (LAP (26), SI): the *Linguistic Atlas of the Mid-Atlantic States* (n=1162 individuals, *LAMSAS* (27)) and the *Linguistic Atlas of the Gulf-States* (n=914 individuals, *LAGS* (28)). The LAP was designed to elicit local and regional variation in the words used for common vocabulary items. For example, the LAP does not investigate lexical variation in the number words, words for days of the week or months of the year, or pronouns for which typically a single word is used in each case. Trained linguist interviewers guided conversations toward pre-determined topics (such as weather, food, buildings, and furniture), recording the words their respondents used for concepts or meanings such as *sofa*, *umbrella*, *chimney*, *canal*, *sit down*, *frost* and *what* (Tables S1,S2).

The LAGS and LAMSAS datasets yielded frequency distributions of the number of speakers per word for 325 and 93 meanings respectively, including meanings such as *cobbler*, *sweet potato*, and *axle* (Figures 1a-c, Tables S1 and S2). Most meanings are nouns (n=301, 72%), followed by verbs (n=53, 12.6%), expressions (n=34, 8.1%), adjectives (n=19, 4.5% and deictics (context-dependent expression, n=11, 2.6%). The number of words reported per meaning is skewed ranging from 2 to 240 with a mean of 30.4±25.3 (median = 25.3, Figure 1d). Because LAP meanings were selected to elicit variation among speakers, this figure over-estimates the average degree of variation in the lexicon, and is probably not representative of what might be thought of as a language's core vocabulary. However, this bias does not affect our study because our interest is in identifying which processes are responsible for different patterns of word-use, and especially cases where a single-word dominates, not the proportion of words explained by *drift*, *directional selection* and *frequency-dependent selection*.

We competed the four models in a Bayesian setting to discover which of the 418 frequency distributions of number of speakers per word (such as Figures 1a-c) they best describe (**Materials and Methods**, SI). Our Bayesian approach yields a posterior probability for each model for each meaning. Because the posterior probabilities sum across models to 1.0 for each meaning, a model's posterior provides a measure of its relative success for that meaning.

Overall, we find little support for random drift (*D*) as a description of the process by which words propagate through a population of speakers (Table 1): some form of selection provides the more probable explanation of the word frequency distributions for over 91% of the meanings, and the results are nearly identical in the two datasets. Drift, or roughly 'say what you hear' or 'copy others' does not provide an adequate description of word-choice. A recent study of three historical grammatical changes also found mixed support for drift (11).

The *FDS+DS* model performs best (Table 1), but appears principally to mimic or compete with *DS* rather than adding a new element to the description of the data: the sum of the *FDS+DS* and *DS* posterior probabilities obtained when all four models are considered (top row, Table 1) correlates across meanings r=0.97 (n=418) with the *DS* posterior probabilities obtained in the absence of *FDS+DS* ('w/o FDS+DS' row, Table 1). We therefore drop *FDS+DS* from further consideration on grounds of parsimony, and analyse the posterior probabilities obtained when we compete the *drift*, *directional selection* and *frequency dependent selection* models.

Our primary interest is in which of the three evolutionary processes (*D*, *DS*, or *FDS*) is most likely to yield strong concordance among speakers as to which word or words to use for a given meaning. In this context, drift tends to provide the best explanation for meanings whose frequency distributions imply the least concordance. For these meanings a variety of words is used by speakers, all co-existing at relatively high frequencies, such as is true of *cobbler* (Figure 1a). Other meanings whose words were governed by drift include *relatives* and *parlor* (Tables S1,S2 and S4).

Where directional selection prevails speakers typically report a smaller number of words, but it is often the case that two or three words are found at relatively high frequencies, with a number of other alternatives at much lower frequencies. Thus, *DS* is the best fitting model for *sweet potato* (Figure 1b) for which both 'sweet potato' and 'yam' were used at high frequencies. *DS* was also the best fitting model for *sofa* – *sofa* and *lounge/couch* used at high frequencies – and *coffin* – *coffin* and *casket* used at high frequencies (Tables S1,S2 and S4). Directional selection, then, yields less variety among speakers than drift but does not seem strong enough in the face of the continual influx of new words to raise one of them to a dominant position.

Where speakers are highly likely to use the same word for a meaning, positive frequency-dependent selection provides the most probable explanation of the word frequencies. This is observed for *axle* (Figure 1c) where one form ('axle') dominates a group of alternatives that only a negligible number of speakers used. Other meanings for which nearly all speakers use the same word and for which *FDS* also provided the best explanation include *towel*, *syrup* and *biscuits* (Tables S1,S2 and S4).

Confirmation that the three different processes yield frequency distributions of word-use with the shapes characteristic of Figures 1a-c can be seen in Figure 2 where the models carve out largely non-overlapping portions of a two-dimensional parameter space defined by two statistics: *2/1 ratio* (the ratio of the 2^{nd} most frequent to the most frequently occurring word) and *heterozygosity* or *H*, a statistic commonly used in genetics to measure the variation in the frequencies of genetic alternatives, here applied to word frequencies (SI). A low 2:1 ratio means that the drop-off in frequency from the most to the second most frequent form is great, and thus is indicative of one word dominating (e.g., *axle* has a low 2:1 ratio). Equally, a low value of *H* also indicates that one word dominates: if one word dominates there is little variation among words in their frequencies – i.e., most respondents use the same word. Both of these features are true of *axle.*

Consistent with these interpretations, frequency dependent selection governs word-choice for meanings that sit in the lower left portion of Figure 2, corresponding to low 2:1 ratio and low heterozygosity. At the other extreme, random drift (*D*) best explains those cases with the least concordance among speakers and consequently they have high *2/1 ratio* and high *heterozygosity* (Figure 2, upper right). Meanings that directional selection explains best tend to fall in the middle.

Where *FDS* is dominant the frequency-dependent selection parameter, *s* (**Methods**; Figure 3 left panel), is more than three times higher than for the remaining meanings (FDS meanings: $\bar{s}$= 0.013±0.014, n=74; D and DS meanings: $\bar{s}$= 0.004±0.002, n=344). *FDS'* posterior probability increases curvilinearly in *s* (Figure 3 right panel), such that when $s \geq 0.006$, *FDS* always provides the best explanation of the data. *FDS* can still predominate even when concordance among speakers appears to be lower (Figure 2, upper right). But these tend to be meanings with two words competing at high frequencies plus an unusually large number of other words at much lower frequencies (*F*-test of *log(no. words*) by winning-model for meanings with 2/1 ratio > 0.5, F=5.12, df=2, p=0.007; all p-values throughout are two-tailed). As a consequence of the large number of words, high levels of *s* (F=4.84, df=2, p<0.007) are required to maintain the two dominant words above the others. For example, for the meaning *a little way*, two phrases – *a little ways* and *a little piece* – were the most commonly used and at nearly equal frequencies.

Characteristics of words are only weakly related to word-use.

We scored all of the words for a representative sample of n=232 meanings (totalling n=252,506 responses, SI, *Word and Meaning Characteristics*) on four attributes related to ease of pronunciation: *complexity* (no. of words in the reply: some replies consist of more than one word, such as *help yourself*), *length* (number of sounds or phones in the reply), number of *obstruent* sounds, corresponding to consonant sounds whose production requires that the airway is obstructed (such as *g* in 'good') and number of *sonorant* sounds or consonants that do not obstruct the airflow. We then correlated words' pronunciation scores with the logarithm of the number of speakers who used them, separately for each meaning. This yielded 232 correlations, each one of which tests the question of whether speakers tend to use the 'better' words. We converted the correlations to z-scores so as to put them on comparable scales and combined them in histograms.

If word characteristics are unrelated to word-choice, we expect the z-score distributions to be centred at zero (corresponding to correlations of zero). Instead, all four distributions are shifted slightly to the left of zero meaning that the words that more of the speakers used have a weak tendency to be easier to pronounce: they are less *complex*, they require fewer sounds (shorter *length*), and they have fewer *obstruents* and *sonorants* (Figure 4 upper row). The effects in the latter three variables might be confounded by *complexity*: replies with more words will have more sounds. However, we find that even after controlling for *complexity* (Figure 4 lower row) the words that are used by more speakers have fewer sounds, including both fewer obstruents and fewer sonorants. Controlling further, for *length*, the effects of obstruents and sonorants disappears (not shown).

The correlations (z-scores) in Figure 4 are small and frequently reversed (any z-score > 0 is opposite to expectation). The weak correlations might reflect the effects of selection itself: by removing 'bad' words the variance among the remaining words in the characteristics related to ease-of-pronunciation is reduced, as is the covariation of these characteristics with the number of speakers who use them. As a consequence, the correlations are unduly influenced by other, background, random factors that affect how many speakers use a word, but which are unrelated to ease-of-pronunciation – an effect consistent with Robertson's secondary theorem(29) from population genetics. Nevertheless, even though small, the correlations in Figure 4 align with the observation from the general lexicon that frequently used words, such as *you, me, he she, I* and the number words tend to be short and easy to pronounce (30), and that languages spontaneously adjust to improve their transmissibility (31). However, we find that the highest frequency words for the meanings the *FDS*, *D* and *DS* models best explain do not differ in their mean scores on the four pronunciation attributes (all p-values >0.18). This suggests that ease of pronunciation of words does not play a strong role in determining the eventual shape of the frequency distributions of numbers of speakers per word.

Characteristics of meanings do not differ among models.

We additionally examined characteristics of the meanings (as opposed to the words). Meanings that drift (*D*) best explained are no more or less likely to be a particular part of speech than expected from the overall data ($p=0.56$), and the same is true of *DS* and *FDS* meanings ($p=0.87$ and $p>0.82$, respectively). We identified for each meaning the word used by the greatest number of speakers, and then obtained the frequency-of-use of that word from the Corpus of Contemporary America Usage, COCA (32). A word's COCA frequency is thus not the same as the number of speakers in our study who used a particular word. Rather, a word's COCA frequency measures how often it appears (relative to words for thousands of other meanings) in a very large sample of word-use (Figure S1). Our interest is to discover whether the top words for the meanings the three models best explained differ in their average COCA frequencies. We find that they do not (geometric mean frequencies in COCA, $p=0.45$): thus it is not the case that, say, words that drift best explained are used less or more often in general, and so on for the other two models. Meanings' mean 'concreteness' scores (33) are also similar among models ($p=0.73$) as are their average ages of acquisition(34) ($p>0.10$). However, among *FDS* meanings the strength of posterior support positively correlates with its concreteness rating ($r=0.38$, $p=0.0004$, n=55), while this relationship is not true of *DS* ($r=0.10$, $p=0.10$, $n=262$) or *D* meanings ($r=-0.16$, $p=0.42$, $n=26$).

Discussion

Our results support Darwin's (3) contention that the words that have survived long enough to become commonplace in everyday speech have got to their positions of favour via a process of natural selection, even if not always by what Müller (2) called their 'inherent virtue'. Thus, the non-selective process of random drift, or roughly 'say what others say', although capable of producing distributions such as those seen in Figures 1a-c, does not provide a general description of word-choice. When new lexical variants are continually being introduced into the vocabulary, as is generally true of language, drift is not strong enough on its own to elevate one or a small number of words to high levels. The answer to the question of how speakers come to use the *same* words, then, is not that they merely copy each other.

Directional selection can to some degree move people towards using the same words. This is seen in the lower *H* scores for words that directional selection explained, and more generally in the weak tendency we observed for speakers to prefer shorter and easier to pronounce words. But this latter effect held across all of the meanings and so does not help to discriminate the meanings that drift and directional best explain from those that frequency-dependent selection explains. Once again, speakers' continual inventiveness with language perhaps removes any simple link between features of words and how often they are currently used: linguistically 'good' words might only have arisen recently and therefore not yet achieved a high frequency, or some otherwise good words might be on their way out of use, having been replaced by others.

By comparison, positive frequency dependence provides an account capable of explaining how speakers come to use the same word for a meaning, such as is typical of what we might think of as the core vocabulary. And this is where we depart from Müller, and we suspect from Darwin, in that under positive FDS a word's 'inherent virtue' seems to play a relatively small role; instead words that, even if from random fluctuations get used at higher frequencies, convert listeners' minds to adopt them as their favoured word, and do so more than would be expected from their frequency alone. This 'conversion' might arise from mere exposure (35) or from active copying of common forms – so-called 'conformist bias'(23).

The value of a conformist bias is perhaps most pronounced in precisely the sort of circumstances that language poses. Communication is important, and so speakers will want to use the right words, but how should they decide which word to use from a number of competing alternatives? In such a situation, an 'agent' that positively 'locked on' to the words that most other people used, or more generally had a motivation to do what most others do, would more quickly achieve a higher or more efficient degree of communication than an agent that merely copied what it heard others saying. Conformity bias such as this has been widely studied in species from fish to humans acting in social and learning milieu where the right course of action is difficult to know (36-38).

Positive frequency dependence also goes some way toward explaining a key puzzle of language, which is how a shared vocabulary can spontaneously self-organise among a group of undirected speakers even when there are potentially many competing alternative words for each meaning. The implicit agreement among speakers that a shared vocabulary requires is made all the more noteworthy by the realisation that, unlike in genetic systems in which there is normally a close connection between a gene's primary sequence and its function (the protein or other product it produces), in language there is seldom a necessary connection between groups of sounds (words) and their meanings, even if some sounds occur more frequently for certain kinds of meanings (39).

But, under positive *FDS*, a word's fitness (likelihood that a speaker will use it as opposed to some other word) continues to increase disproportionately as it becomes more common, and this force eventually propels the word to fixation, that is, it becomes the sole word used for that meaning. Unlike with drift or directional selection, this increasing strength of selection continues in spite of the constant influx of new words, which by virtue of being at low frequencies will have low fitnesses. Indeed, at fixation the force of positive *FDS* is greatest and so positive frequency dependence could also help to explain how some words can remain paired with a meaning for hundreds or even thousands of years (7, 40), far exceeding the time-span of the possibly three to four generations that might separate the oldest and youngest speakers in a group (frequently used forms are also less buffeted by the effects of drift (11, 12)).

Meanings for which respondents collectively reported a large number of alternative words (e.g., Figure 1a,b), could still be cases of frequency-dependent selection acting independently within sub-contexts of that meaning, each of which has its own favoured word or set of words. This might be true of meanings that admit a wider coverage or breadth of contexts than others. The meaning 'cobbler' for example, best explained by drift, might include a wider range of contexts than the meaning 'axle', best explained by *FDS*. If the words corresponding to the various sub-contexts of meanings with greater breadth are combined into a single distribution, something like that of Figure 1a (cobbler) could emerge, but be hiding sub-contexts in which a single word or small number of words dominates. We have no evidence that this is the case, but if true some of our drift or directional selection meanings might actually be FDS meanings.

Our modelling assumes that the number of different word forms for a meaning is in a stochastic equilibrium fluctuating around some average maintained by the loss of existing words and the gain of new ones. This is, of course, an approximation, but consistent with this assumption, the number of different words per meaning correlates 0.87 for the sixty-six meanings that occur in both the LAMSAS and LAGS datasets, and the top two words for many of the meanings are the same (SI). Nevertheless, it is possible that some of our word distributions in which two or a variety of words is commonly used could eventually resolve to a single dominant word, or in other cases a contender to a dominant word might arise. Our modelling also treats each respondent as having just a single word for each meaning, when in fact most respondents would probably recognise all or nearly all of the various words that other respondents reported. Our assumption is that respondents are telling us the word they would be most likely to use.

It does not escape our attention that the mechanism of frequency-dependent selection is also the mechanism that would govern most fads or the rapid spread of novel cultural forms and ideas. In this sense, language is laid bare as a cultural phenomenon, subject at least in part to fluctuations in usage that could often be little more than whimsy in origin. And, indeed, such linguistic-fads are seen, as in the rapid spread of slang and other vernacular elements. Why the core lexicon is relatively shielded from the ephemeral existence of most fads is an intriguing subject for lexicographers, linguists, sociologists and others interested in cultural change. One possibility is that most language-use is designed to convey factual information while fads are at least partly driven by status and identity signalling that derives it force from novelty and thereby loses momentum as a phenomenon becomes common; and this might give insight into what constitutes a mere fad versus something that will become more lasting.

Materials and Methods

Models. We suppose that the number of speakers who use each of the *i= 1….k* different words for a particular meaning (e.g., Figures 1a-c) represents the long-term outcome of a mutation-selection balance process in which new words or expressions continually arise at some rate θ and are continually affected by selection.

Let

$$W_i = \frac{x_i^{(1+s)} w_i}{\sum_i (x_i^{(1+s)} w_i)} \qquad (1)$$

where, x_i is the frequency of speakers in a population who use alternative form *i* (*i=1…k*), *s* represents the strength of frequency-dependent selection acting on *i* ($s \geq 0$), w_i is a coefficient denoting the intrinsic fitness of word *i* independently of how many speakers use it, and the summation in the denominator is over all forms *i*. Defined this way, W_i is the expected frequency in the next generation of word *i* relative to the other words for a particular meaning.

When *s=0* and all w_i = 1, all words are equivalent and equation 1 describes *random drift (D)*. Drift (D) supposes that a number of neutral alternative words exist for a given meaning, that new forms are continually introduced and that speakers use words in proportion to the number of other speakers who use them.

Setting *s=0* but allowing w_i to vary among words, yields a model of *directional selection* (*DS*) that incorporates drift but allows some words to be better or worse than others by an amount that depends upon the magnitude of w_i. The w_i are not optimised or fit to the observed frequencies as this would assume that 'better' words have higher frequencies. Rather, they are assigned to words at random as they enter the lexicon (see *Model estimation*, below).

If $s > 0$, but all $w_i = 1$, equation 1 describes *positive frequency-dependent selection (FDS)*. Under *positive FDS* the likelihood that a speaker will use a word increases disproportionately to the number of other speakers using it. The strength of frequency dependence is characterised by the parameter s (equation 1), where positive frequency dependence corresponds to $s > 0$. Finally, we created a model that combines *positive-frequency dependent selection with directional selection (FDS+DS)*.

Model estimation. We used Approximate Bayesian Computation (ABC, (41, 42), SI) to estimate models' abilities to predict each meaning's frequency distribution of speakers. ABC is widely-used in population studies because it can incorporate the effects of drift and selection acting within populations. ABC simulates models a large number of times with parameters drawn randomly from prior distributions, retaining the simulations closest to the observations. These retained runs sample from the posterior distribution of model parameters most likely to have given rise to an observed set of data, y.

The ABC design is (41):

i) Draw $\Theta_i \sim \pi(\Theta)$,
ii) Simulate $x_i \sim p(x \mid \Theta_i)$.
iii) Reject Θ_i if $x_i \neq y$, where y are the observed data. The subset of draws from Θ that produce x_i similar to y define the posterior distribution of Θ, $p(\Theta \mid y)$.

Here, Θ is a vector corresponding to the parameters of the evolutionary model, $\pi(\Theta)$ is the prior distribution of Θ (see SI), and the x_i are simulated from this prior. Alternative forms of the vector Θ define the drift, directional selection and frequency-dependent selection models, according to Equation 1. The acceptance/rejection at step iii is achieved by use of a set of summary statistics defined on the data (SI).

Simulations (step ii) of the directional selection model randomly associate the w_i terms with a word when it enters the lexicon reflecting the possibility that, for example, a word newly entering the lexicon, and thus at low frequency, might nevertheless have $w_i > 1$. The prior distribution of these weights is centred at 1.0 and then falls away in both directions in a manner roughly corresponding to exponential decline following Ohta (43). The weights then influence, along with the effects of drift, how the word spreads through the population of speakers over generations of word transmission. For description of the priors on the other parameters see the SI.

Our simulations presume a genealogical process (from the perspective of the word) in which words move from speaker to speaker with one of three outcomes: the word might remain unchanged, it can mutate to a new form, or an existing word can replace the word another speaker uses. Over the long term this leads to an equilibrium distribution of word frequencies that is governed by the forces of drift and selection as represented in each model. Word frequencies vary from one generation to the next because fitter forms are more likely to be copied, or because a speaker's word might be replaced by another 'fitter' word, or by mutation creating a new word

Model Comparisons. A model's performance relative to the other models is assessed by its Bayesian posterior probability, given by

$P(M_i|D) = \frac{P(D|M_i)p(M_i)}{\sum_i P(D|M_i)p(M_i)}$ where $P(M_i|D)$ is the probability of the data under model *i*, and $p(M_i)$ is the prior probability of model *i*. $P(D|M_i)$ is calculated as the proportion of simulations in which model *i* best describes the summary statistics. A model's posterior probability is proportional to the number of simulations (out of a large number) for which the model best matched the *S(y)*. We then record the 'winner' for each meaning as the model with the highest posterior probability.

Availability of code. Code to implement the models is available from A.M. and M.B.

Linguistic Atlas Project Data. All raw data are available via the Linguistic Atlas Project websites and handbooks. See SI, *Materials and Methods*. In addition, we make available all of our files and filtering criteria available at the Open Science Framework (https://osf.io), public project *MotherTongue*.

Meaning characteristics

COCA word frequencies. We identified the most commonly reported word given for each of the meanings in our sample. We then consulted the Corpus of Contemporary American English (COCA(32)), and recorded that word's frequency-of-appearance (written and spoken use), noting its rank-order position in the list.

Concreteness scores. We obtained 'concreteness' rankings for 40,000 commonly used English words and two-word expressions(33), where concreteness was defined as the extent to which the meaning refers to something that can be experienced directly through the senses (1-5 scale where 5 is concrete and 1 is abstract). We found matches or near-matches in this list to the highest frequency word for n=292 of the meanings in our sample of n=418. The concreteness scores correlate r=0.94 with concreteness ratings obtained from an earlier study of 4291 words(44).

Age of Acquisition. We recorded the mean age of acquisition(34) for each of our meanings. We found, as above, matches or near matches to n=312 of our meanings.

Acknowledgements

This work was supported by Advanced Investigator Award 268744 'MotherTongue' to MP from the European Research Council. MB acknowledges NERC grant NE/K006088/1, AM acknowledges BBSRC grant BB/L018594/1, and AC acknowledges the New Zealand Royal Society Catalyst Seeding Fund. We thank Kathryn Harris for her help with collecting 'concreteness' scores and filtering some of the response data. We thank Bill Kretzschmar for allowing us to make the raw data available.

References and Notes

1. Schleicher A (1869) *Darwinism Tested by the Science of Language* (John Camden Hotten, London).
2. Müller M (1870) The Science of Language. *Nature* 1:256-259.
3. Darwin CR (1871) *The Descent of Man and Selection in Relation to Sex* (John Murray, London).
4. Blevins J (2004) *Evolutionary phonology: The emergence of sound patterns* (Cambridge University Press).
5. Croft W (2000) *Explaining language change: an evolutionary approach* (Pearson Education).
6. Labov W (2011) *Principles of linguistic change, cognitive and cultural factors* (John Wiley & Sons).
7. Pagel M, Atkinson QD, & Meade A (2007) Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* 449(7163):717-720.
8. Lieberman E, Michel J-B, Jackson J, Tang T, & Nowak MA (2007) Quantifying the evolutionary dynamics of language. *Nature* 449(7163):713-716.
9. Kimura M (1984) *The neutral theory of molecular evolution* (Cambridge University Press).
10. Wright S (1931) Evolution in Mendelian populations. *Genetics* 16(2):97-159.
11. Newberry MG, Ahern CA, Clark R, & Plotkin JB (2017) Detecting evolutionary forces in language change. *Nature* 551:223-226.
12. Reali F & Griffiths TL (2010) Words as alleles: connecting language evolution with Bayesian learners to models of genetic drift. *Proceedings of the Royal Society of London B: Biological Sciences* 277(1680):429-436.
13. Kretzschmar W (2015) *Language and Complex Systems* (Cambridge University Press, Cambridge).
14. Bentley RA (2008) Random drift versus selection in academic vocabulary: An evolutionary analysis of published keywords. *PloS one* 3(8):e3057.
15. Hahn MW & Bentley RA (2003) Drift as a mechanism for cultural change: an example from baby names. *Proceedings of the Royal Society of London B: Biological Sciences* 270(Suppl 1):S120-S123.
16. Bentley RA, Hahn MW, & Shennan SJ (2004) Random drift and culture change. *Proceedings of the Royal Society of London B: Biological Sciences* 271(1547):1443-1450.
17. Ewens WJ (1972) The sampling theory of selectively neutral alleles. *Theoretical population biology* 3(1):87-112.
18. Zipf GK (1949) *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology* (Addison-Wesley Press).
19. Blythe RA & Croft W (2012) S-curves and the mechanisms of propagation in language change. *Language* 88(2):269-304.
20. Acerbi A & Bentley RA (2014) Biases in cultural transmission shape the turnover of popular traits. *Evolution and Human Behavior* 35(3):228-236.
21. Chouteau M, Arias M, & Joron M (2016) Warning signals are under positive frequency-dependent selection in nature. *Proceedings of the National Academy of Sciences* 113(8):2164-2169.
22. Simon HA (1955) On a class of skew distribution functions. *Biometrika* 42(3/4):425-440.
23. Boyd R & Richerson P (1985) *Culture and the evolutionary process* (University of Chicago Press, Chicago, IL).
24. Mesoudi A & Lycett SJ (2009) Random copying, frequency-dependent copying and culture change. *Evolution and Human Behavior* 30(1):41-48.

25. Henrich J (2001) Cultural transmission and the diffusion of innovations: Adoption dynamics indicate that biased cultural transmission is the predominate force in behavioral change. *American Anthropologist* 103(4):992-1013.

26. Davis AL (1969) A Compilation of the Work Sheets of the Linguistic Atlas of the United States and Canada and Associated Projects.

27. Kretzschmar WA (1993 (2007)) *Handbook of the linguistic atlas of the Middle and South Atlantic States* (University of Chicago Press).

28. Pederson L, McDaniel SL, Bailey G, & Bassett M (1986) *Linguistic Atlas of the Gulf states, vol. 1: Handbook for the Linguistic Atlas of the Gulf States* (Athens: University of Georgia Press).

29. Robertson A (1968) The spectrum of genetic variation. *Population biology and evolution*:5-16.

30. Zipf GK (1949) Human Behaviour and the Principle of Least-Effort. Cambridge MA edn. (Addison-Wesley, Reading).

31. Kirby S, Cornish H, & Smith K (2008) Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences* 105(31):10681-10686.

32. Davies M (2009) The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics* 14(2):159-190.

33. Brysbaert M, Warriner AB, & Kuperman V (2014) Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods* 46(3):904-911.

34. Kuperman V, Stadthagen-Gonzalez H, & Brysbaert M (2012) Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods* 44(4):978-990.

35. Zajonc RB (1968) Attitudinal effects of mere exposure. *Journal of personality and social psychology* 9(2p2):1.

36. Henrich J & McElreath R (2003) The evolution of cultural evolution. *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews* 12(3):123-135.

37. Kendal RL, Coolen I, & Laland KN (2004) The role of conformity in foraging when personal and social information conflict. *Behavioral Ecology* 15(2):269-277.

38. Whiten A, Horner V, & De Waal FB (2005) Conformity to cultural norms of tool use in chimpanzees. *Nature* 437(7059):737.

39. Blasi DE, Wichmann S, Hammarström H, Stadler PF, & Christiansen MH (2016) Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*:201605782.

40. Pagel M, Atkinson QD, Calude AS, & Meade A (2013) Ultraconserved words point to deep language ancestry across Eurasia. *Proceedings of the National Academy of Sciences* 110(21):8471-8476.

41. Beaumont MA (2010) Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution and Systematics* 41(379-406):1.

42. Beaumont MA & Rannala B (2004) The Bayesian revolution in genetics. *Nature Reviews Genetics* 5(4):251-261.

43. Ohta T (1977) Extension to the neutral mutation random drift hypothesis. *Molecular evolution and polymorphism. National Institute of Genetics, Mishima, Japan*:148-167.

44. Coltheart M (1981) The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology* 33(4):497-505.

Acknowledgements This work was supported by Advanced Investigator Award 268744 'MotherTongue' to MP from the European Research Council. MB acknowledges NERC grant

NE/K006088/1, AM acknowledges BBSRC grant BB/L018594/1, and AC acknowledges the New Zealand Royal Society Catalyst Seeding Fund. We thank Kathryn Harris for her help with collecting 'concreteness' scores and filtering some of the response data. We thank Bill Kretzschmar for allowing us to make the raw data available.

Author contributions All authors contributed to all aspects of the research.

Correspondence and requests for materials should be addressed to M.P.: m.pagel@reading.ac.uk

Conflicts of interest The authors declare that they do not have any conflicts of interest.

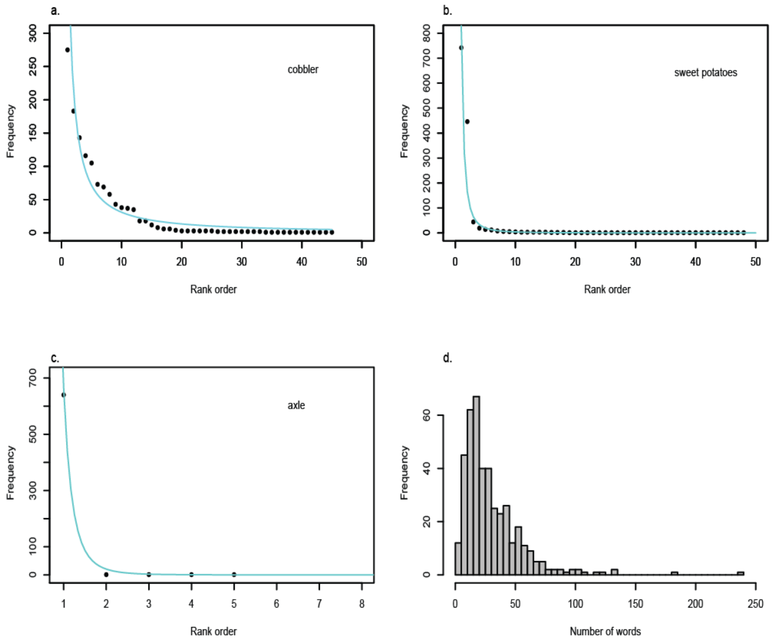

Figure 1. panels a,b,c) The frequencies of alternative words (y-axis) plotted against their rank order (x-axis), with smooth curve of the form $y=ax^{-b}$ fitted for descriptive purposes. The exponent b increases (steeper drop-off) from panel a to c reflecting the decreasing frequency of the second word relative to the first (note: attenuated x-axis of panel 1c disguises the steepness of the exponent b). Panel d) Frequency distribution of the number of words per meaning for the n=418 meanings (mean=30.4±25.3, median = 25.3).

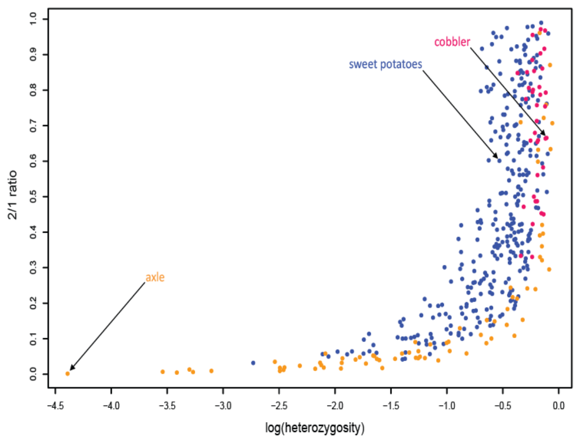

Figure 2. Ratio of the frequency of the 2nd most commonly used word for a meaning to the highest frequency word (2/1 ratio) plotted against the logarithm of heterozygosity (H), showing regions where each of the models performs best: mustard = frequency-dependent-selection (FDS), blue = Directional Selection (DS), magenta = Drift (D). Lower 2/1 ratios indicate high agreement among speakers (greater dominance of first word as in Figure 1c). Heterozygosity (H, see text and SI for definition) varies between 0 and 1 and measures the degree to which word frequencies are uniform (high H, indicating low degree of concordance among speakers) or are concentrated in one or a small number of words (low H, high concordance among speakers). FDS explains word frequencies characterised by high concordance among speakers (low 2/1 ratio and low H), or relatively low 2/1 ratio (low for any given level of heterozygosity). DS explains intermediate levels of both measures. Drift (D) best characterises meanings with a variety of words at relatively high frequencies (low concordance among speakers). Mean 2/1 ratios and mean heterozygosity differ significantly among the models such that D>DS>FDS (all p-values < 0.001).

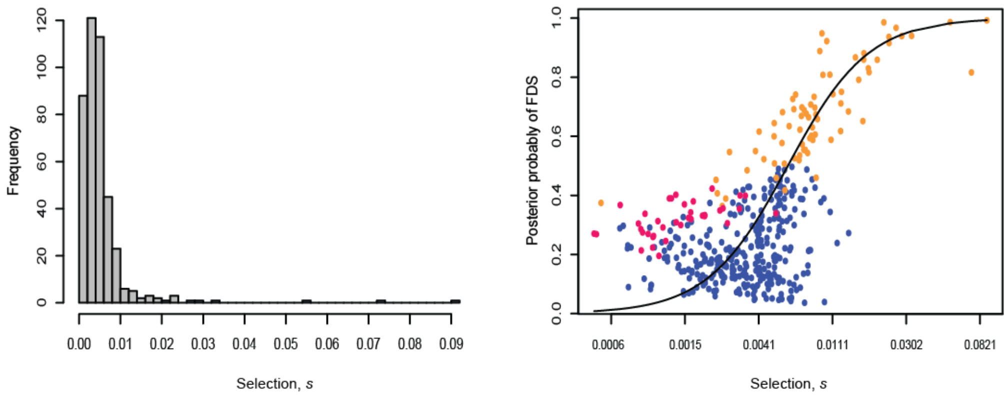

Figure 3. Left panel: frequency distribution of selection coefficients, s: mean = 0.005±0.007 (median = 0.004). Right panel: Posterior probability of FDS model against the size of the selection coefficient showing curvilinear relationship (note: x-axis on log-scale). Points are all FDS posterior probabilities but colour-coded to indicate the model that had the highest posterior probability for that meaning: mustard = FDS, blue = DS, magenta = D. The value of s for which FDS' posterior probability > 0.5 corresponds to s~0.006.

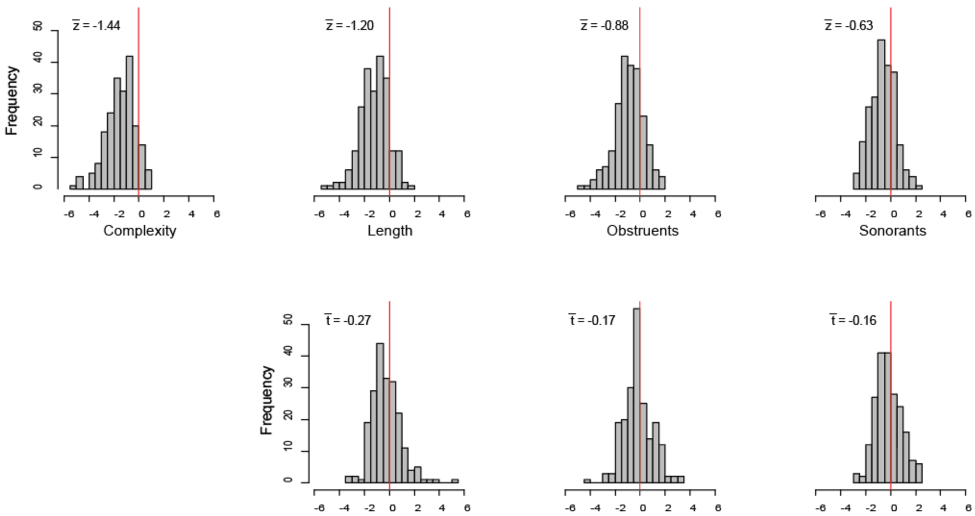

Figure 4. Upper panel: histograms of z-transformed rank order correlations between an attribute score and word frequency for four attributes related to ease-of-pronunication: complexity, length, obstruents and sonorants (n=232 meanings, SM). Histograms are shifted to negative side of 0.0 indicating that words used more often for a given meaning tend to score lower (better) on the attribute (see text)., All z-scores, $p < 1^{-10}$. Lower panel: histograms of t-scores after controlling for complexity (responses with more words have more sounds). Length remains significant (p<0.002), while effect sizes are small (average t=-0.21) for obstruents (p=0.06) and sonorants (p=0.04). Controlling for length, the effect of obstruents and sonorants disappears.

Table 1. Percentage of 'winners' by model, and their summary statistics

Dataset	D	FDS	DS	FDS+DS
Full dataset (n=418 meanings)	**8.6**	**16.3**	**35.6**	**39.5**
LAGS, n=325	8.6	16.9	35.7	38.8
LAMSAS, n=93	8.6	14.0	35.5	41.9
Full dataset (w/o FDS+DS)	**8.8**	**17.7**	**73.4**	
Statistic (mean±s.e.m)				
2/1 ratio	0.70±0.03	0.18±0.03	0.45±0.02	
Heterozygosity, *H*	0.82±0.01	0.42±0.04	0.58±0.01	
Example meanings (Tables S1, S2 and S4)	cobbler, parlor, hay shed, relatives	axle, towel, biscuits, syrup	sweet potatoes, sofa, coffin, skunk	

Upper section: Percentage of n=418 meanings where the model shown has the highest posterior probability (Methods): D=drift, FDS=frequency-dependent selection, DS=directional selection, FDS+DS=combined FDS and DS model (text); Lower section: means of two key summary statistics (text) for cases where the model shown above has highest posterior probability.