Drift as a driver of language change: An artificial language experiment

Rafael Ventura[1], Joshua B. Plotkin[2], and Gareth Roberts*[3]

[1]Social and Cultural Evolution Working Group, University of Pennsylvania
[2]Department of Biology, University of Pennsylvania
[3]Department of Linguistics, University of Pennsylvania

Abstract

Over half a century ago, George Zipf observed that less frequent words tend to have entered the language more recently. Since then, corpus studies have accumulated evidence that rare words have higher rates of replacement and regularisation. Two main hypotheses have been proposed to explain this pattern: (a) less frequent words change more because deterministic selection against innovation is weaker for such words, and (b) less frequent words change more because stochastic drift is stronger in smaller populations. Here, we report an experimental test of these hypotheses. Participants were tasked with learning a miniature language consisting of two nouns and two plural markers. Nouns occurred at different frequencies and were subjected to treatments that varied drift and selection. Using a model that accounts for participant heterogeneity, we measured the rate of noun regularisation and the strength of selection and drift in participant responses. Our results indicate that the elevated rate of regularisation we observed in the low-frequency nouns was attributable to drift alone. These results add to a growing body of evidence that drift may be a major driver of language change.

1 Introduction

George Zipf noted a series of statistical regularities in natural languages [1]. Best known among these is that word frequency is inversely proportional to word rank [2]. A similar regularity is that frequent words tend to be shorter than rare ones [3, 4]. But Zipf also observed that less frequent words are more likely to be recent borrowings or coinages [1]. Recent work has found evidence for this, reporting that frequent words tend to have lower rates of replacement [5]. A related finding is that frequent words also have lower rates of

*Corresponding Author: gareth.roberts@ling.upenn.edu regularisation. In a study of a corpus spanning over 1,200 years, rare past-tense verbs in English were found to regularise more rapidly than frequent ones [6]. In a study of Google Books and Twitter data, infrequent English verbs were similarly found to regularise more often [7].

It is not clear, however, why a negative correlation between frequency of use and regularisation and replacement rates should hold. Selection might drive this pattern [5]. If, for instance, communicative pressures act more strongly on more frequent terms, selection against innovation could act in a frequency-dependent manner to stabilise linguistic forms. This is a plausible hypothesis but concrete evidence for or against it is lacking. Another possibility is that the pattern is simply driven by drift, with infrequent words being replaced and regularised at higher rates by chance as sampling variance is greater in smaller population sizes. In a neutral model of word replacement, drift alone was shown to give rise to the negative correlation between frequency and replacement rate [8]. In a corpus of contemporary American English, smaller effective population sizes were inferred for infrequent words [9], which suggests that drift may be sufficient to explain why rare past-tense verbs in English regularise faster.

There are thus different potential explanations for the observed correlation between frequency of use and replacement and regularisation rates. As existing empirical studies are based on corpus data, they cannot conclusively answer this question: Corpora allow us to study language in highly realistic conditions but cannot track the entire trajectory of a language, nor control the many different factors that affect language change [10]. More to the point, questions about the strength of drift and selection—whether in cultural or biological evolution—hit methodological challenges of their own. Disentangling the effects of drift and selection in molecular variation has been one of the primary challenges for 20th-century evolutionary biology [11]. In the case of language change, one method used to infer the strength of drift and selection was shown to be sensitive to choices of data binning [12]. In particular, a test for selection versus a null hypothesis of neutral drift was shown to depend on how a corpus is parsed into time intervals (bins of equal amount of data versus equal duration of time) to extract time-series. Slightly different results were obtained in a binary classification of drift versus selection when analysing the same corpus with a deep neural network [13], suggesting that a better approach is to quantify the strength of selection and the strength of drift [9]. This is the approach we take here, quantifying both selection and drift to help understand how word frequency is related to rates of replacement and regularisation.

To address this question, we conducted an artificial-language experiment. Such experiments have already revealed several factors, including learning, age, memory, and multigenerational transmission, that up- and down-regulate regularisation [14, 15, 16, 17, 18, 19, 20]. Data from similar experiments have been used to investigate the role of drift and selection in the emergence of simple communication systems [21]. But no experiment to date has investigated whether it is drift or selection that gives rise to the negative correlation between frequency of use and regularisation.

Here, we take drift to be any source of unbiased stochasticity in the acquisition, processing, and production of language. Several studies have already detected signatures of drift in cultural evolution, including language change [22, 23, 9]. A similar phenomenon could be at the root of frequency-dependent replacement and regularisation rates. If language users tend to acquire, recall, and produce linguistic forms based on a sample of forms they encounter, there will be stochastic variation in the frequency of alternative forms simply because language users sample a finite set of language-related stimuli. The strength of this stochasticity may be greater for some words than for others. For example, suppose that English speakers choose the past-tense forms “strived” or “strove” for the verb “strive” in proportion to how often they encounter each variant. Speakers might then tend to learn the regular or irregular form simply because chance exposes them to one variant more often than the other. As stochasticity due to sampling is greater for smaller samples, variation may be lost more rapidly in rare verbs as a result. In this way, drift could drive the correlation between frequency of use and rates of replacement and regularisation.

Selection, on the other hand, is any directional bias in the acquisition, recollection, and production of language. Several studies have found evidence for selection driving language change [24, 25, 26]. Selection could likewise be responsible for frequency-dependent rates of replacement and regularisation. For instance, regularised alternatives to high-frequency forms (e.g., “goed” vs. “went”) are likely to incur higher processing costs relative to the established irregular form. Cognitive biases of this kind may be stronger for more frequent words, thereby inhibiting regularisation in a frequency-dependent manner. The negative correlation between frequency of use and rates of replacement and regularisation could thus be driven by differential selection strengths.

To study the role that drift and selection play in shaping the negative correlation between frequency of use and regularisation, we conducted an experiment in which participants were tasked with learning a miniature artificial language. The language consisted of two nouns and two plural markers. Nouns occurred at different frequencies and were subjected to drift and selection of varying strengths. By measuring their regularisation, we were then able to determine whether low-frequency nouns did in fact regularise more than high-frequency nouns and, if so, whether this was due to drift or selection. Our experimental setup therefore allowed us to test three main hypotheses: Hypothesis 1 was that the greater regularisation of low-frequency words results from stronger drift on low-frequency terms, Hypothesis 2 was that the greater regularisation of low-frequency words results from stronger selection on high-frequency terms, and Hypothesis 3 was that the greater regularisation of low-frequency words results from both effects.

With this study we thus seek to further our understanding of the cultural evolutionary forces that shape human languages. We also hope to contribute to the experimental study of cultural evolution and strengthen the link between evolutionary theory and the experimental study of language.

2 Method

Our experiment was pre-registered with the Open Science Foundation (https://osf.io/72kqa/?view_only=d0f2776850fe4c5e970b68237049c400).

2.1 Participants

We recruited 400 participants (189 female; 200 male; five non-binary; one other; six did not report their gender) through Prolific. Of these 180 reported being between 18 and 30 years old, 92 reported being between 30 and 40, 78 reported being between 40 and 50, 38 reported being between 50 and 60, 10 reported being 60 years old or older, and two did not report their age. To be eligible, participants had to report English as their native language. Participants were informed that this was a study on an alien language and were asked to give their consent before taking part in the experiment. Participants were paid a base rate of $1.00 for participating in the study. Participants were also told that they would receive a 50% bonus depending on the accuracy of their answers; in reality, however, all participants who completed the study were given the 50% bonus. Data from 10 participants who were more than two standard deviations in either direction from the mean completion time were discarded.

2.2 Alien Language

Participants were trained on a miniature artificial language composed of nouns for two different referents embedded in an English sentence. To facilitate learning, each noun was composed of a root consisting of two CV syllables. The root consonants were identical to that of the English words with the same meaning ("buko" for book and "hudo" for hand). For each root, participants were asked to learn a singular and a plural form. The singular form consisted in the unmarked root; the plural was formed by adding a suffixed marker to the root with two possible variants, "-fip" and "-tay" [20]. Markers were randomly assigned to roots between participants. Each noun belonged to one of two frequency classes. The low-frequency noun was shown six times during the training and the testing phase; the high-frequency noun was shown 18 times. Nouns were randomly assigned to a frequency class.

2.3 Procedure

Participants interacted with a custom-made website programmed with PennController for Ibex [27], an online experiment scripting tool, and hosted on the PCIbex Farm (`expt.pcibex.net`). Instructions were provided on screen before each stage of the experiment. The experiment began with a training phase in which participants were asked to learn an alien language (henceforth the *input language*). The training phase was followed by a testing phase in which participants were asked to use the language (henceforth the *output language*). The training phase consisted of two subphases, so participants passed through the following phases:

1. *Training phase*

   (a) *Noun Training*
   Participants were first shown a picture depicting a single object. Below the image,

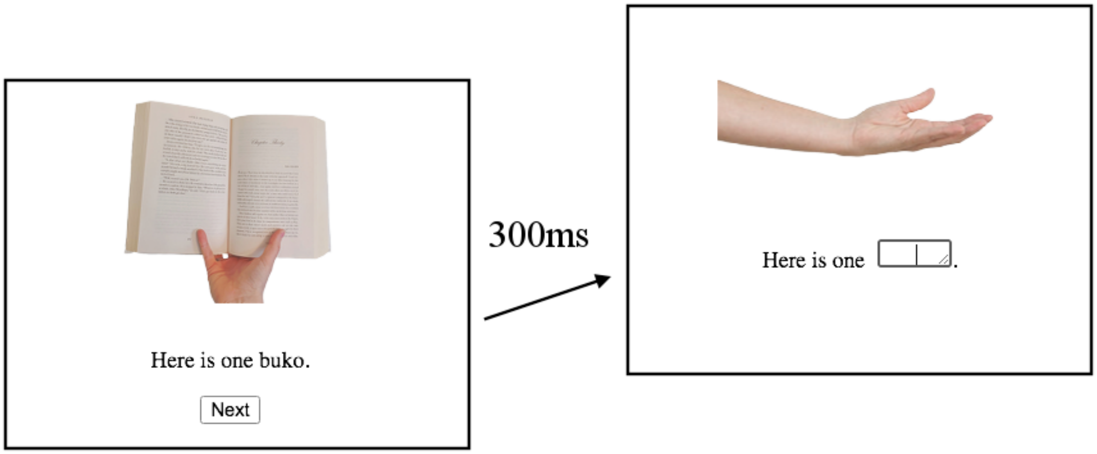
a caption with the sentence "Here is one buko" or "Here is one hudo" instructed participants on the singular noun used to refer to the object. After clicking a *Next* button, participants were shown another image depicting another object. Each picture was shown once in random order, with a 300 ms mask between trials. Participants were then shown the same pictures two more times, alternating between a trial in which they were again shown an object with a full sentence referring to it and a trial in which they were shown an object and asked to complete a sentence of the form "Here is one ____". Participants had to enter the correct form of the noun to move on to the next trial. If the form was correct, participants were told so; if the form was incorrect, a box popped up reminding them of the correct form and asking them to try again. Participants were therefore shown each object-noun pair twice and asked to enter the noun corresponding to each object once.

(b) *Plural Training*

Participants were shown a picture depicting three partly overlapping instances of the same object (Figure 1). The objects were the same as the ones shown during noun training. After clicking a "Next" button, participants were shown another image. Below each image, a caption with the sentence "Here are several buko+MARKER" or "Here are several hudo+MARKER" instructed participants on the plural noun used to refer to the objects. Depending on frequency class, each picture was shown either six or 18 times. Pictures were shown in random order, with a 300 ms mask between trials. At random intervals, participants were also shown images containing a single object and asked to type the correct noun to complete the sentence. Whenever an incorrect form was entered, a box again reminded the participants of the correct form and asked them to try again. Single objects appeared only once.

2. *Testing phase*

This phase was similar to plural training. Participants were again shown pictures depicting three partly overlapping instances of the same objects and with the same frequency as in the plural training phase. At random intervals, they were also shown pictures of single objects. During this phase, however, participants were asked to type the corresponding noun in each trial to complete the sentence. Participants therefore had to enter either the singular or the plural form of the noun, depending on the picture shown. In the plural case, participants were told that the form was correct provided that it was seven characters long and that it contained one of the two plural markers at the end. If it was incorrect, participants were asked to try again without being told what the correct form was. In the singular case, participants were told that the form was correct provided that their answer was four characters long. Otherwise, participants were asked to try again without being told what the correct form was.


Here is one buko.

Here is one .

Figure 1: **Training and Testing.**

2.4 Conditions

There were two conditions. In the Drift Condition, all nouns in the input language occurred with plural markers at a 1:1 ratio. During plural training, the low-frequency noun therefore occurred with one plural marker in three trials and with the other plural marker in the remaining three trials. Similarly, the high-frequency noun occurred with one plural marker in nine trials and with the other plural marker in the remaining nine trials. The purpose of the Drift Condition was to establish an input language in which there was no directional pressure to regularise forms. That is, if the language changed, it would be as a result of drift rather than selection.

In the Selection Condition, all nouns in the input language occurred with plural markers at a 5:1 ratio. During plural training, the low-frequency noun therefore occurred with one plural marker in five trials and with the other plural marker in the remaining trial. Similarly, the high-frequency noun occurred with one plural marker in 15 trials and with the other plural marker in the remaining three trials. Low- and high-frequency nouns differed with respect to what plural marker was more common. For example, if the low-frequency noun occurred five times with the plural marker "-fip" and only once with the marker "-tay", then the high-frequency noun occurred 15 times with the plural marker "-tay" and three times with the marker "-fip". The purpose of the Selection Condition was to establish an input language in which there was an asymmetry between plural markers and, as a consequence, the potential for a directional pressure—that is, selection—to regularise forms. In particular, we predicted that nouns would lose the secondary marker and adopt the primary one with a probability greater than would occur by chance alone.

In the Selection Condition, the more common plural marker was designated as the "primary" marker and the less common plural maker the "secondary" marker. To allow for comparison across conditions, we arbitrarily labelled half of the markers as "primary" and the other half as "secondary" in the Drift Condition as well. In both conditions, nouns that occur at least once with both markers were designated as the"irregular" nouns; nouns that occur with a single plural marker exclusively were termed "regular". In the input language for both conditions, all nouns were therefore irregular. In the output language, nouns could be either regular or irregular depending on the behaviour of the participant.

We defined a Regularisation Index (RI) as the change in the proportion of irregular nouns between the input and output languages [6]. As there was only one low-frequency and one high-frequency noun, this meant that the RI value for each frequency class for a given participant could take a value of either 0 (meaning that the noun was irregular in the output language) or 1 (meaning that it was regular in the output language). To validate results based on the RI, we consider the mean change in conditional entropy—another commonly used measure of regularisation (see Supplementary Material A).

The RI offers a natural way to formulate our hypotheses. If the greater regularisation of low-frequency nouns is due to drift alone (Hypothesis 1), then the RI is higher in the low-frequency class than in the high-frequency class in both the Drift and the Selection Conditions. If the greater regularisation of low-frequency nouns is due to selection alone (Hypothesis 2), then the RI is higher in the low-frequency class than in the high-frequency class only in the Selection Condition. If the greater regularisation of low-frequency nouns is due to both drift and selection (Hypothesis 3), then the difference in RI between the low- and the high-frequency class is higher in the Selection than in the Drift Condition.

2.5 Preliminary Study

To test the viability of our experimental design, we conducted a preliminary study using the methods as described so far. The experiment was pre-registered with the Open Science Foundation (https://osf.io/ryc3j/?view_only=0026aa2c67184e339d2d7d478a8024ac). Results from this preliminary study are described in full in Supplementary Material B. In particular, results revealed that the pool of participants was heterogeneous with respect to the experimental task. In the Selection Condition, the distribution of marker counts had a single peak and a long tail: most participants chose the secondary marker with probability equal to or less than its initial frequency, but many also randomised their choice of markers (Supplementary Material, Figure 2 *right*). At the same time, the distribution of marker counts was trimodal in the Drift Condition: most participants randomised their choice of markers, but some chose either one of the markers exclusively (Supplementary Material, Figure 2 *left*). Results from this preliminary study informed the statistical analysis plan for the current experiment, as reported below.

2.6 Statistical Analysis

To test our hypotheses, we used a binomial logistic model with noun regularity (i.e., regular or irregular) as the dependent dichotomous variable and frequency (i.e., low or high frequency) and selection (i.e., presence or absence) as independent dichotomous variables. In particular, the model took the following form:

$$ln\left(\frac{p}{1-p}\right) = b_0 + b_1F + b_2S + b_3FS \tag{1}$$

where p is the proportion of regular nouns, F is a variable indicating frequency class (low: 0; high: 1), S is a variable indicating absence or presence of selection (absence: 0; presence: 1), and FS is a term representing the interaction between frequency class and selection. In this model, b_1 measures the main effect of frequency class, b_2 measures the main effect of selection, and b_3 measures the interaction effect of frequency class and selection on noun regularisation. Hence, if b_1 is significantly less than zero but b_3 is not, the model supports the hypothesis that the greater regularisation of low-frequency terms is due to weaker drift in the high frequency class (Hypothesis 1); if b_3 is significantly less than zero but b_1 is not, the model supports the hypothesis that the greater regularisation of low-frequency terms is due to stronger selection in the high frequency class (Hypothesis 2); and if both b_1 and b_3 are significantly less than zero, the model supports the hypothesis that the greater regularisation of low-frequency terms is due to weaker drift and stronger selection in the high-frequency class (Hypothesis 3).

We then conducted a manipulation check to confirm the presence of selection for the primary marker in the Selection Condition. As our preliminary study revealed a heterogeneous participant pool, we first built a model representing different types of participants. The model assigns probability q that participants choose a single plural marker regardless of input language ("simplifiers"). Further, the model assigns probability r that participants choose plural markers according to a binomial distribution with parameters n and 0.5, where n is the population size and 0.5 means that participants randomise their choice of plural markers ("randomisers"). Finally, the model assigns probability $1 - q - r$ that participants choose plural markers according to the Wright-Fisher model with selection ("regularisers").

The Wright-Fisher model, commonly used in evolutionary biology and shown to be equivalent to models of iterated learning [8], represents change in a population of two types as a draw from a binomial distribution with parameters n and $f(n, s)$, where n is the population size in the Wright-Fisher model and $f(n, s)$ is given by:

$$f(n, s) = \frac{i \cdot e^s}{i \cdot e^s + (n - i)} \quad , \tag{2}$$

where i is again the number of markers in the input language, n is the number of nouns in the given frequency class, and s is the selection coefficient (for further details of the Wright-Fisher model, see Supplementary Material C).

Estimates of q, r, and s for our experimental sample were calculated by selecting the parameter values that maximise the sum of the log-likelihoods of the data. In particular, the estimate $\hat{s}$ was given by the following expression:

$$(\hat{q}, \hat{r}, \hat{s}) = \underset{q,r\in\Delta, s\in[-5,5]}{\text{argmax}} \sum_{j=1}^{N} log\Bigg(r \cdot P(i_j|j = 1) + q \cdot P(i_j|j = 2) + (1 - r - q) \cdot P(i_j|j = 3)\Bigg) \tag{3}$$

where $j = 1$ if participant j is a regulariser, $j = 2$ if j is a simplifier, and $j = 3$ if j is a randomiser. Here, Δ denotes the simplex volume $\{(q, r) \in [0, 1]^2 \mid q + r \leq 1\}$. In this way, we were able to simultaneously estimate the selection coefficient $\hat{s}$ among regularisers and the composition $(\hat{q}, \hat{r})$ of the participant pool.

The two-tailed 95% confidence interval for $\hat{s}$ was given by the log-likelihood ratio and thus included all values of s satisfying $\ell(s) - \ell(\hat{s}) \leq 1.92$, where $\ell(s)$ is the sum of log-likelihood given s, maximising over parameters (q, r). Similarly, the two-tailed 95% confidence regions for $(\hat{q},\ \hat{r})$ included all values of (q, r) satisfying $\ell(q, r) - \ell(\hat{q}, \hat{r}) \leq 2.99$, where $\ell(q, r)$ is the sum of log-likelihood given (q, r), maximising over the parameter s.

3 Results

Analysis was conducted using Python [28] and Julia [29]. Data and scripts for the experiment are available at `https://osf.io/5m9ak/?view_only=aaa9660774964e76838903533f6f0c16`.

Regularisation was higher for low- than for high-frequency nouns in both conditions (Figure 2). In particular, RI estimates for low- and high-frequency were equal to 0.51 ± 0.07 and 0.48 ± 0.07 in the Drift Condition ($N = 194$) and equal to 0.75 ± 0.06 and 0.73 ± 0.06 in the Selection Condition ($N = 196$).

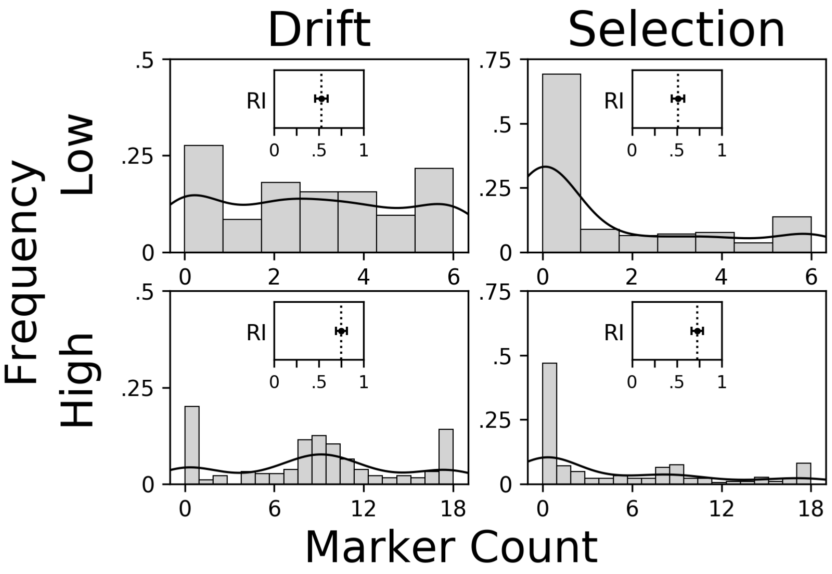


Figure 2: **Marker Counts and Regularisation Index (RI).** Distribution of irregular marker counts (empirical values shown by grey bars; density plot given by black line). Regularisation Index (RI) is the change in the proportion of regular nouns between input and output languages (insets; error bars show 95% confidence interval. Drift: $N = 194$. Selection: $N = 196$

We then conducted a manipulation check to ensure that there was selection against the secondary marker in the Selection Condition and no selection the Drift Condition. As expected, we found evidence of selection for the primary marker in the Selection Condition: among regularisers, $\hat{s}$ was equal to $-2.3 \pm (0.9, 0.6)$ and $-2.1 \pm (0.3, 0.4)$ for low- and high-frequency nouns (Figure 3). Estimates for low- and high-frequency nouns were similar in value, indicating that selection for the primary marker was of roughly the same strength in both frequency classes.

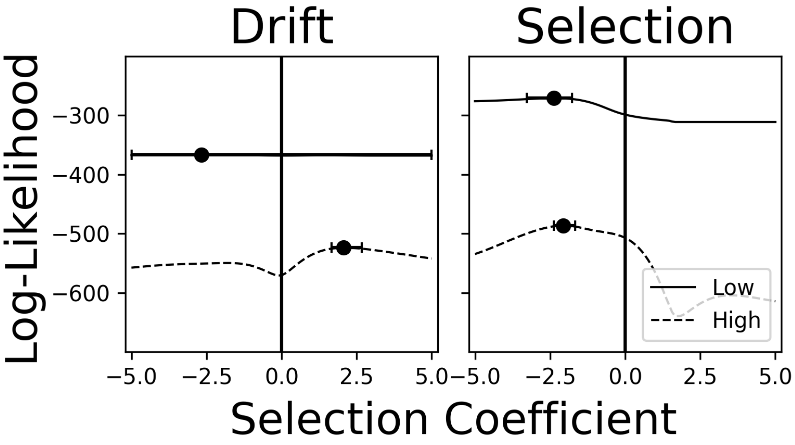


Figure 3: **Sum of log-likelihoods.** Curves show the sum of log-likelihoods for the data given the selection coefficient for regularisers in the population model; circles show the maximum-likelihood estimates of the selection coefficient; and error bars show two-tailed 95% confidence intervals.

In the Drift Condition, there was no evidence of selection among regularisers in the low-frequency class: $\hat{s}$ was $-2.1 \pm (2.93, 7.1)$, with the 95% confidence interval spanning the entire range of selection coefficients sampled (Figure 3). Contrary to our expectation, however, our estimate for the selection coefficient was positive in the high-frequency class: $\hat{s}$ was $1.97 \pm (0.4, 0.71)$. This might seem like an indication that there was selection in the Drift Condition, which would clash with a central assumption of our analysis plan.

But this was likely not the case. In the Drift Condition, estimates for the proportion of randomisers in the low- and high-frequency classes were 0.56 and 0.6 (Figure 4). Similarly, estimates for the proportion of simplifiers were 0.37 and 0.31. Both participant types therefore made up almost the entirety of the participants, with regularisers comprising only 0.07 and 0.09 in the low- and high-frequency classes. It is thus likely that our maximum-likelihood algorithm detected positive selection among regularisers in the high-frequency class simply due to noise in the data, as we estimate there are very few regularisers in our sample (namely, $0.09 \times 193 \approx 17$), and the chi-squared asymptotic confidence interval on ML estimates is known to be a poor approximation in small samples. The fact that negative selection was detected in our preliminary study corroborates this point (see Supplementary Material B, Figure 1).

In the Selection Condition, on the other hand, our estimates show that regularisers made

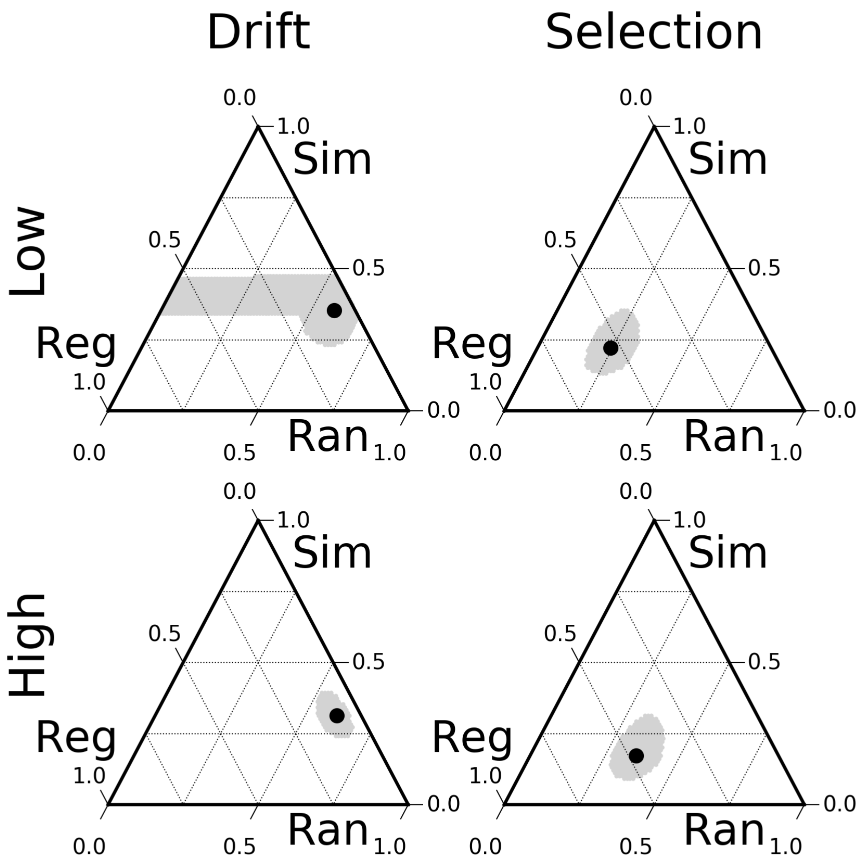


Figure 4: **Population Composition.** Black dots show the maximum-likelihood value of the population composition in the population model with proportion p of randomisers, proportion q of simplifiers, and proportion $1 - p - q$ of regularisers; areas in grey show 95% confidence regions.

up 0.54 and 0.48 of the population in the low- and high-frequency classes; the proportion of randomisers was 0.24 and 0.35, and the proportion of simplifiers was 0.2 and 0.17. Together with the finding that selection was negative among regularisers in both frequency classes of the Selection Condition, our manipulation check therefore suggests that selection for the primary marker was indeed present in the Selection Condition but likely absent in the Drift Condition.

Since regularisation was higher among low-frequency nouns in both conditions and drift is expected to be stronger at low frequencies, our results provide evidence in support of the hypothesis that low-frequency nouns regularise more because of drift. Moreover, selection coefficients were approximately equal for low- and high-frequency nouns in the Selection Condition, meaning that the regularisation pattern in this condition could not have been due to selection. Results therefore support the hypothesis that drift alone drives the regularisation pattern.

Our binomial logistic regression model provides further support for this hypothesis. Frequency class had a negative effect on noun regularity, with high-frequency nouns being significantly more likely to regularise in both conditions ($b_1 = -0.42 \pm 0.21$; $p = 0.047$; Table 1). As expected, selection had a large positive effect on noun regularity ($b_2 = 1.2 \pm 0.21$; $p < 0.0001$). At the same time, there was no detectable effect of the interaction between frequency class and selection ($b_3 = -0.39 \pm 0.29$; $p = 0.19$). Given that frequency class had a significant effect on regularisation but the interaction between frequency class and selection did not, our results provide support for the hypothesis that the greater regularisation of low-frequency nouns was driven by stronger drift in the low-frequency class. It is important to keep in mind, however, that our regression model does not take into account the population composition.

Table 1: **Logistic regression model.** The model was given by $ln(\frac{p}{1-p}) = b_0 + b_1 F + b_2 S + b_3 FS$. Significant results at the 0.05 level are marked with '*'.

	β	SE	p
intercept (b_0)	-0.31	0.14	0.032*
frequency (b_1)	-0.42	0.21	0.047*
selection (b_2)	1.2	0.21	< 0.0001*
freq. × sel. (b_3)	-0.39	0.29	0.19

We applied the same analysis to the data from our preliminary study (Section 2.5), which yielded consistent results (see Supplementary Material B for full results).

4 Discussion

We conducted an experiment in which participants were exposed to a small miniature language consisting of two nouns and two plural markers. We manipulated both the frequency of the nouns and the presence of drift and selection. We found a difference in regularisation between low- and high-frequency nouns, and our analysis indicated that this was due to drift alone. Although our study is based on small-scale laboratory experiments, the results are consistent with the view that drift plays a large role in the negative correlation between frequency of word use and rates of regularisation and replacement that Zipf first noted in natural languages [1] . Our study therefore adds to a growing body of evidence suggesting that drift is a major driver of language change, including patterns of regularisation.

In addition, our results highlights the risk of assuming—rather than showing—that participants approach any experimental task as a single homogeneous population. Indeed, the participant pool was far from homogeneous with respect to the experimental task in our study. While most participants behaved as expected in regularising the use of plural markers in the Selection Condition, a significant portion of participants did not. Instead, many opted either to randomise their choice of markers or to simplify the task by using a single marker throughout the experiment—in both cases effectively disregarding the initial marker frequencies. In the Drift Condition, most participants behaved as expected and randomised their choice of markers. But a non-negligible portion of the participants also simplified the task by using a single marker.

There are, however, some limitations to discuss. For one, the difference in noun frequency alone was insufficient to produce a detectable difference in selection strength across the two frequency classes of the Selection Condition. As a result, selection could not be responsible for the difference in regularisation across frequency classes, although this might well be the case in natural systems. In natural languages, for example, there may be stronger selection against replacement and regularisation for high-frequency words if these words function as anchors during language acquisition and learning [30]. It is also possible that other factors, such as morpheme length or phonological complexity, influence the strength of selection, further complicating the picture.

Another limitation of our study is that participants learned but did not *use* the language to communicate with other participants. Yet the social context in which natural languages are used can also gives rise to selection. Social meaning and identity influences which linguistic forms are used [31, 32], and communicative interaction shapes the cultural evolution of language [33, 34]. Even if drift is the primary mechanism of regularisation, as our results suggest, its role may therefore be modulated by selection in different contexts.

A further limitation concerns the size of the artificial language. As it consisted of two nouns and two affixes, it was the smallest possible language for our purposes. This was done to maintain careful control over how the language was learned: participants were likely to learn the morphemes of the language, preventing differences in learning success from constituting a nuisance variable and thereby enhancing ecological validity. It also made the experiment short and quick to run, which allowed us to gather a large sample cost effectively. The negative flip side, of course, was that the language differed even more from natural languages than is typically the case and increased the potential for demand characteristics to play a role. But the downside of such a simple language was outweighed by its benefits in our experiment, where tight control over marker and noun frequencies was important.

There are a number of clear avenues for future research. To our knowledge this study is the first experimental treatment of Zipf's observation about regularisation and frequency of use. Future work might concentrate on employing more complex languages, or on incorporating more complex social contexts, including direct communication between participants [34, 32] or simulated communication [35]. Quite a number of language-internal and -external factors may well play important roles in modulating—or even supplanting—the role of drift. Future work could, for instance, implement selection of different strengths across different frequency classes, or compare different sources of selection [21].

Finally, we consider the integration of natural-language observation, experimental linguistic data, and mathematical models from biology to be a strength of our approach, and are pleased that this has been typical of related research [36, 9]. We feel that this kind of interdisciplinary approach is a very positive sign for the future of the language sciences [37].

References

[1] Zipf GK. Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Cambridge: Addison-Wisley; 1949.

[2] Piantadosi ST. Zipf's word frequency law in natural language: A critical review and future directions. Psychonomic bulletin & review. 2014;21(5):1112–1130.

[3] Kanwal J, Smith K, Culbertson J, Kirby S. Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. Cognition. 2017;165:45–52.

[4] Sigurd B, Eeg-Olofsson M, Van Weijer J. Word length, sentence length and frequency–Zipf revisited. Studia Linguistica. 2004;58(1):37–52.

[5] Pagel M, Atkinson QD, Meade A. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. Nature. 2007;449(7163):717–720.

[6] Lieberman E, Michel JB, Jackson J, Tang T, Nowak MA. Quantifying the evolutionary dynamics of language. Nature. 2007;449(7163):713–716.

[7] Gray TJ, Reagan AJ, Dodds PS, Danforth CM. English verb regularization in books and tweets. PloS one. 2018;13(12):e0209651.

[8] Reali F, Griffiths TL. Words as alleles: Connecting language evolution with Bayesian learners to models of genetic drift. Proceedings of the Royal Society B: Biological Sciences. 2010;277(1680):429–436.

[9] Newberry MG, Ahern CA, Clark R, Plotkin JB. Detecting evolutionary forces in language change. Nature. 2017;551(7679):223.

[10] Galantucci B, Garrod S, Roberts G. Experimental semiotics. Language and Linguistics Compass. 2012;6(8):477–493.

[11] Kimura M. Evolutionary rate at the molecular level. Nature. 1968;217(5129):624–626.

[12] Karjus A, Blythe RA, Kirby S, Smith K. Challenges in detecting evolutionary forces in language change using diachronic corpora. arXiv preprint arXiv:181101275. 2018.

[13] Karsdorp F, Manjavacas E, Fonteyn L, Kestemont M. Classifying Evolutionary Forces in Language Change Using Neural Networks. Evolutionary Human Sciences. 2020:1–40.

[14] Kirby S, Tamariz M, Cornish H, Smith K. Compression and Communication in the Cultural Evolution of Linguistic Structure. Cognition. 2015;141:87–102.

[15] Hudson Kam CL, Newport EL. Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. Language Learning and Development. 2005;1(2):151–195.

[16] Samara A, Smith K, Brown H, Wonnacott E. Acquiring variation in an artificial language: Children and adults are sensitive to socially conditioned linguistic variation. Cognitive Psychology. 2017;94:85–114.

[17] Perfors A. When do memory limitations lead to regularization? An experimental and computational investigation. Journal of Memory and Language. 2012;67(4):486–506.

[18] Ferdinand V, Kirby S, Smith K. The cognitive roots of regularization in language. Cognition. 2019;184:53–68.

[19] Reali F, Griffiths TL. The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. Cognition. 2009;111(3):317–328.

[20] Smith K, Wonnacott E. Eliminating unpredictable variation through iterated learning. Cognition. 2010;116(3):444–449.

[21] Tamariz M, Ellison TM, Barr DJ, Fay N. Cultural selection drives the evolution of human communication systems. Proceedings of the Royal Society B: Biological Sciences. 2014;281(1788):20140488.

[22] Bentley RA. Random drift versus selection in academic vocabulary: An evolutionary analysis of published keywords. PloS one. 2008;3(8).

[23] Hahn MW, Bentley RA. Drift as a mechanism for cultural change: an example from baby names. Proceedings of the Royal Society of London Series B: Biological Sciences. 2003;270:S120–S123.

[24] Amato R, Lacasa L, Díaz-Guilera A, Baronchelli A. The dynamics of norm change in the cultural evolution of language. Proceedings of the National Academy of Sciences. 2018;115(33):8260–8265.

[25] Sindi SS, Dale R. Culturomics as a data playground for tests of selection: Mathematical approaches to detecting selection in word use. Journal of Theoretical Biology. 2016;405:140–149.

[26] Stadler K, Blythe RA, Smith K, Kirby S. Momentum in language change: a model of self-actuating s-shaped curves. Language Dynamics and Change. 2016;6(2):171–198.

[27] Zehr J, Schwarz F. PennController for Internet Based Experiments (IBEX); 2018.

[28] Van Rossum G, Drake Jr FL. Python Reference Manual. Centrum voor Wiskunde en Informatica Amsterdam; 1995.

[29] Bezanson J, Edelman A, Karpinski S, Shah VB. Julia: A Fresh Approach to Numerical Computing. SIAM review. 2017;59(1):65–98. Available from: https://doi.org/10.1137/141000671.

[30] Frost RLA, Monaghan P, Christiansen MH. Mark my Words: High Frequency Marker Words Impact Early Stages of Language Learning. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2019;45(10):1883.

[31] Roberts G, Fedzechkina M. Social Biases Modulate the Loss of Redundant Forms in the Cultural Evolution of Language. Cognition. 2018;171:194–201.

[32] Sneller B, Roberts G. Why Some Behaviors Spread while Others Don't: A Laboratory Simulation of Dialect Contact. Cognition. 2018;170:298–311.

[33] Galantucci B. Experimental semiotics: A new approach for studying communication as a form of joint action. Topics in Cognitive Science. 2009;1(2):393–410.

[34] Wade L, Roberts G. Linguistic Convergence to Observed Versus Expected Behavior in an Alien-Language Map Task. Cognitive Science. 2020;44(4):e12829.

[35] Buz E, Tanenhaus MK, Jaeger TF. Dynamically Adapted Context-specific Hyper-articulation: Feedback from Interlocutors Affects Speakers' Subsequent Pronunciations. Journal of Memory and Language. 2016;89:68–86.

[36] Karjus A, Blythe RA, Kirby S, Smith K. Quantifying the Dynamics of Topical Fluctuations in Language. Language Dynamics and Change. 2020;10:86–125.

[37] Roberts G, Sneller B. Empirical Foundations for an Integrated Study of Language Evolution. Language Dynamics and Change. 2020;10(2):188–229.