Title

Regularization of nouns due to drift, not selection: An artificial-language experiment

Permalink

https://escholarship.org/uc/item/6x57h78b

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 43(43)

ISSN

1069-7977

Authors

Ventura, Rafael
Plotkin, Joshua
Roberts, Gareth

Publication Date

2021

Peer reviewed

Regularization of nouns due to drift, not selection: An artificial-language experiment

Rafael Ventura (rhtventura@gmail.com)
Social and Cultural Evolution Working Group, University of Pennsylvania

Joshua B. Plotkin (jplotkin@sas.upenn.edu)
Department of Biology, University of Pennsylvania

Gareth Roberts (gareth.roberts@ling.upenn.edu)
Department of Linguistics, University of Pennsylvania

Abstract

Corpus data suggests that frequent words have lower rates of replacement and regularization. It is not clear, however, whether this holds due to stronger selection against innovation among high-frequency words or due to weaker drift at high frequencies. Here, we report two experiments designed to probe this question. Participants were tasked with learning a simple miniature language consisting of two nouns and two plural markers. After exposing plural markers to drift and selection of varying strengths, we tracked noun regularization. Regularization was greater for low- than for high-frequency nouns, with no detectable effect of selection. Our results therefore suggest that lower rates of regularization of more frequent words may be due to drift alone.

Keywords: Zipf; language change; artificial-language experiment; selection; drift; cultural evolution; language evolution

Over 70 years ago, Zipf (1949, p. 116) observed that less frequent words are more likely to be recent borrowings or coinages. Some recent studies have also found evidence that frequently occurring words tend to have lower replacement or regularization rates (Pagel, Atkinson, & Meade, 2007; Lieberman, Michel, Jackson, Tang, & Nowak, 2007; Gray, Reagan, Dodds, & Danforth, 2018). It is not clear, however, why such a relationship should hold. Pagel et al. (2007) speculated that cultural selection against regularization and replacement might be stronger on high-frequency words, thereby driving the pattern. Another possibility is that the pattern is simply driven by drift, with infrequent words having higher rates of replacement and regularization due to sampling error, which is greater at lower frequencies (Reali & Griffiths, 2010; Newberry, Ahern, Clark, & Plotkin, 2017).

There are, to our knowledge, no experimental studies on this question. The existing empirical studies are based on corpus data, which provide high ecological validity but do not allow us to track the entire trajectory of a language or control the factors involved in change. They also run up against methodological challenges, being sensitive to such factors as data binning (Karjus, Blythe, Kirby, & Smith, 2018; Karsdorp, Manjavacas, Fonteyn, & Kestemont, 2020).

We conducted a preregistered experiment to investigate whether the negative correlation between frequency and regularization might be due to drift (understood as non-directional bias in acquisition, processing, and production of language) or selection (understood as directional bias). Based on the results of this experiment, we conducted a preregistered replication with an identical design but an adjusted analysis plan.

In both experiments, participants were tasked with learning a miniature artificial language that consisted of two nouns and two plural markers. To implement drift of different strengths, we varied noun frequency; to implement selection of different strengths, we varied the frequency of one plural ending relative to the other. We then measured noun regularization as the fraction of nouns that came to be used with one ending only. This allowed us to test three main hypotheses: that greater regularization of low-frequency words results from stronger selection on high-frequency terms (Hypothesis 1), from stronger drift on low-frequency terms (Hypothesis 2), or from both (Hypothesis 3).

Experiment 1

Experiment 1 was pre-registered (`https://osf.io/ryc3j`).

Method

Participants We recruited 400 native-English-speaking participants (207 female; 183 male; four non-binary; five chose not to report their gender) through Prolific (`www.prolific.co`). 324 reported being 18–40 years old, 71 reported being 40 years old or older, and five chose not to report their age. Participants were paid $1.00 for participating. As motivation they were also told that they would receive a 50% bonus based on the accuracy of their answers; in reality, all participants who completed the study were given the bonus and thus received $1.50 in total.

Artificial Language To ensure learnability we constructed a miniature artificial language of the smallest size needed to test our hypotheses. The language consisted of two nouns and two plural endings. Each noun referred to one of two different referents (hand and book). Word forms similar to their English counterparts were chosen to facilitate learning: "hudo" meaning *hand* and "buko" meaning *book*, and words were presented embedded in English sentences. Participants were exposed to a singular and a plural form for each noun. The singular consisted in the unmarked root; the plural was formed by adding a suffixed marker to the root with two possible variants, "-fip" and "-tay" (cf. Smith & Wonnacott, 2010). Plural markers were randomly assigned to roots between participants. Nouns were randomly assigned to a frequency class.

Procedure Experiment software was created using PennController for Ibex (Zehr & Schwarz, 2018) and hosted on the PCIbex Farm (`expt.pcibex.net`). The experiment began with a *training phase* in which participants were exposed to the artificial language (henceforth the *input language*). Then, in the *testing phase*, participants were asked to produce the language (henceforth *output language*).

The training phase consisted of two subphases. In *noun training* trials participants saw a picture of a single object with the caption "Here is one NOUN" (where NOUN was the target noun). Participants could click *Next* to advance to the next trial. Each picture was shown once in random order, with a 300 ms break between trials. Participants were then shown the same pictures twice more, alternating between a trial in which they were again shown an object with a sentence caption 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 noun to proceed. Noun training was followed by *plural training*, which resembled noun training except that each image had three overlapping instances of the same object and the caption text read "Here are several NOUN+MARKER". Depending on frequency class, each picture was shown either six or 18 times. At random intervals, participants were also shown images of a single object and asked to provide the correct noun.

The testing phase was similar to plural training. Participants were shown pictures depicting three instances of the same object, each with the same frequency as in the previous phase. At random intervals, they were shown pictures of single objects. Now, however, they were asked to type the corresponding noun to complete the sentence. In the case of plurals, participants were told that their form was correct if it was seven characters long and contained one of the plural markers at the end. Otherwise, they were asked to try again. In the singular case, participants were told that any four-character form was correct. Otherwise, they were asked to try again.

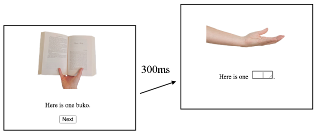

Figure 1: **Training and Testing.**

Conditions There were two conditions. In the *Drift* Condition, nouns occurred with both plural markers at a 1:1 ratio; in the *Selection* Condition, nouns occurred with plural markers at a 5:1 ratio (see Table 1). But low- and high-frequency nouns differed with respect to which marker was more common. For example, if the low-frequency noun occurred more often with "-fip", the high-frequency noun occurred more often with "-tay".

Table 1: Drift and Selection Conditions. Indicated are the number of trials for the two nouns (N_1 and N_2) and the two plural makers (M_1 and M_2).

Drift	M_1	M_2	Total
N_1	3	3	6
N_2	9	9	18

Sel.	M_1	M_2	Total
N_1	1	5	6
N_2	15	3	18

In the Drift Condition, we expected no directional pressure for regularization in favor of one or the other plural marker. Regularization could therefore be due to drift, but not selection. In the Selection Condition, we expected directional pressure for regularization in favor of the more common marker. Since drift is also present in any finite number of trials, regularization could be due to drift or selection.

In the Selection Condition, we call the more common marker for each noun the "primary" marker and the less common marker the "secondary" marker for that noun. To facilitate comparison across conditions, we arbitrarily labeled half of the markers as "primary" and the other half as "secondary" for each noun in the Drift Condition as well. In both conditions, nouns that occur at least once with both markers are termed "irregular"; nouns that occur only with a single marker are termed "regular". The binary coding of nouns may appear to be too stringent but it was necessary for our binomial logistic regression model, described below.

Statistical Analysis Following Lieberman et al. (2007), we defined a Regularization Index (RI) as the change in the proportion of irregular nouns between the input and output languages. We used RI to make simple comparisons of regularization across conditions: If Hypothesis 1 is correct, then RI should be higher in the low- than the high-frequency class only in the Selection Condition; if Hypothesis 2 is correct, then RI should be higher in the low-frequency class in both conditions; and if Hypothesis 3 is correct, then the difference in RI between the low- and the high-frequency classes should be greater in the Selection than in the Drift Condition.

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

$$ln\left(\frac{p}{1-p}\right) = b_0 + b_1 I(f) + b_2 I(s) \qquad (1)$$

where p is the proportion of regular nouns, $I(f)$ indicates drift (low: 0; high: 1), and $I(s)$ indicates selection (absence: 0; presence: 1).

As a manipulation check, we used a Wright-Fisher model with selection to represent the change between input and output languages (cf. Reali & Griffiths, 2010). The Wright-Fisher model represents change in a population of two individual types as a draw from a binomial distribution with parameters n and $f(n,s)$, where n is the population size and $f(n,s)$ is given by:

$$f(n,s) = \frac{i(1+s)}{i(1+s)+(n-i)} \tag{2}$$

where i is the number of individuals of a particular type and s is the selection coefficient.

In our experiment, a Wright-Fisher population corresponds to the ensemble of noun tokens in a given frequency class; the individual types correspond to the different plural markers that nouns can take. Accordingly, $n = 6$ in the low-frequency class and $n = 18$ in the high-frequency class. We took the focal type to be the secondary marker.

To estimate the selection coefficient against the secondary marker, we computed the likelihood of transitions from the input language to every possible output language given different values of s. The maximum-likelihood estimate of $\hat{s}$ is then given by the value of s that maximizes the sum of the log-likelihoods for all participants. In other words, $\hat{s}$ is given by the following expression:

$$\hat{s} = \underset{s \in [-1,1]}{\operatorname{argmax}} \sum_{j=1}^{N} log\left(P\left(i_j | Bin\big(n, f(n,s)\big)\right)\right) \tag{3}$$

where i_j is the secondary marker count in the output of participant j, $P(i_j|Bin(n, f(n,s)))$ is the likelihood of i_j given the Wright-Fisher model, and the sum is over all participants. Two-tailed 95% confidence intervals were given by $\ell(s) - \ell(\hat{s}) \leq 1.92$, where $\ell(s)$ is the sum of log-likelihoods given s.

A positive estimate would indicate that selection favored the secondary marker, while a negative estimate would indicate selection against it. An estimate of 0 would indicate the absence of selection.

Results of Experiment 1

Mean completion time in minutes was 8.9 ($s.d. = 5.2$) and 8.6 ($s.d. = 4.6$) for the Drift and Selection Conditions. Data from 12 participants whose completion time was more than two standard deviations from the mean were excluded.

Regularization as measured by RI was higher for low- than for high-frequency nouns in both the Drift ($N = 193$) and the Selection conditions ($N = 195$), with RI estimates for low- and high-frequency equal to 0.51 ± 0.07 and 0.46 ± 0.07 in the Drift Condition and equal to 0.79 ± 0.06 and 0.79 ± 0.06 in the Selection Condition (Figure 2). We then estimated the strength of selection using our maximum-likelihood algorithm. Surprisingly, our estimates of selection for both frequency classes in the Selection condition had roughly the same positive value: $\hat{s}$ was equal to $0.33 \pm (0.16, 0.2)$ and $0.35 \pm (0.1, 0.1)$ for low- and high-frequency nouns (Figure 3). This was surprising because a positive value indicates selection in favor of the secondary marker, contrary to a central assumption of our experiment and analysis plan.

Our regression model indicates that frequency class had a negative effect on noun regularity (-0.65 ± 0.15; $p < 0.0001$; Table 2). In contrast, selection had a positive effect on noun regularity (0.98 ± 0.15; $p < 0.0001$). However, the presence of selection for the secondary marker makes it difficult to analyze these results according to our original analysis plan.

Table 2: Logit regression model: $ln(\frac{p}{1-p}) = b_0 + b_1 I(f) + b_2 I(s)$; see Method for variable definitions. Significant results at the 0.05 level are marked with '*'.

	β	SE	p
intercept (b_0)	-0.083878	0.126405	0.5070
frequency (b_1)	-0.651267	0.150075	< 0.0001*
selection (b_2)	0.981622	0.150078	< 0.0001*

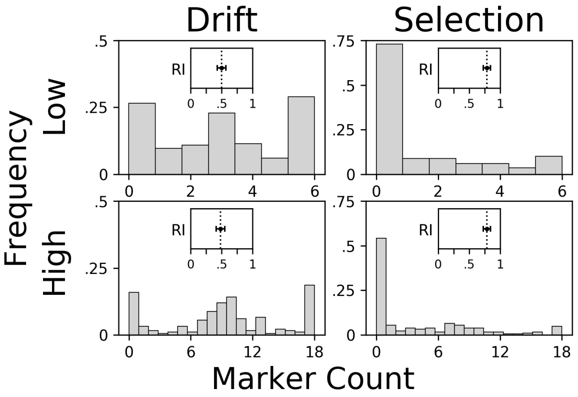

Figure 2: **Marker Counts and Regularization Index (RI) for Experiment 1.** Frequency of irregular marker counts. *Insets:* mean change in proportion of regular nouns between input and output languages (RI) with 95% confidence interval. Drift: $N = 193$. Selection: $N = 195$.

Discussion of Experiment 1

Closer inspection of the data suggested that selection for the secondary marker may be an artifact of the learning task. As Figure 2 shows, the distribution of marker counts had a single peak and a long tail in the Selection Condition. This suggests that most participants chose the secondary marker with probability equal to or less than its initial frequency but that many also randomized their choice of markers. When participants choose markers at random, the frequency of the secondary marker increases. As our algorithm was designed to detect selection alone, an increase in the frequency of the secondary marker could only be interpreted as positive selection. In the Drift Condition, on the other hand, the distribution of marker counts was trimodal: most participants randomized their choice of markers (central mode) but some chose either one of the markers exclusively (left and right modes).

We therefore sought to account for these findings with a single model, which worked as follows. In both conditions, we assume that proportion r of participants chooses the primary marker according to the Wright-Fisher model. We also assume that the population has proportion q of "simplifiers"

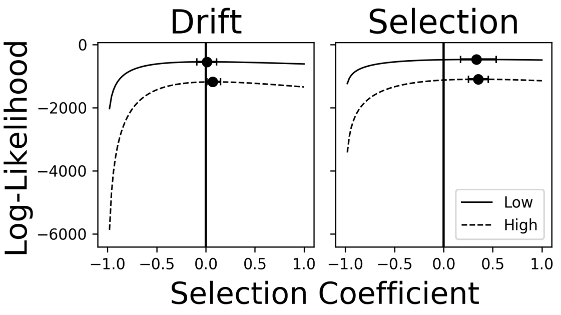

Figure 3: **Sum of log-likelihoods for Experiment 1.** Curves show sum of log-likelihoods given selection coefficient for regularizers in population model; error bars show maximum-likelihood values with 95% confidence intervals, indicating selection for the secondary marker.

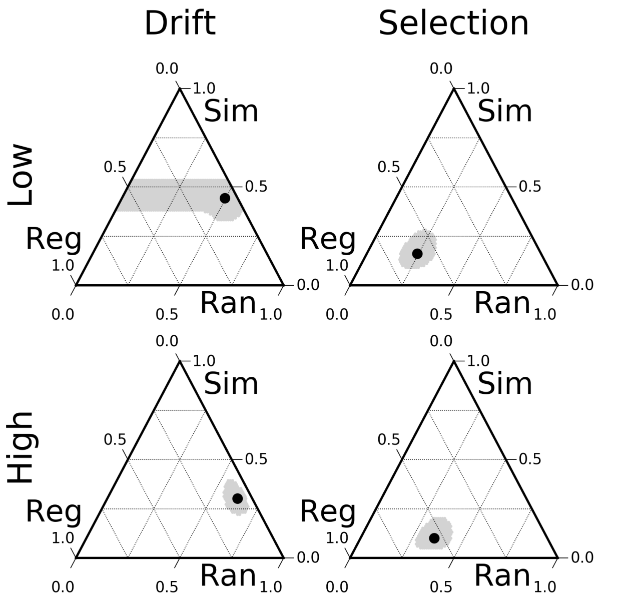

Figure 4: **Population Composition for Experiment 1.** Filled black circles show maximum-likelihood composition in population model with proportion p, q, and $1-p-q$ of randomizers, simplifiers, and regularizers, indicating that the participant pool was heterogeneous and variable across conditions (areas show 95% confidence regions).

who always choose a single marker and proportion $1-r-q$ of "randomizers" who randomize their choice of marker.

The model fits the data for Experiment 1 well (Figure 4). But as this model was not included in our analysis plan, we designed a second experiment taking the model into account. Experiment 2 was therefore designed to replicate the main finding of Experiment 1 that regularization is higher for low- than for high-frequency nouns, and to account for the possibility that the behavior of different participant types might interfere with our algorithm's ability to detect selection. We also expanded the logistic model to include an interaction term between frequency class and presence of selection. We describe Experiment 2 in the next section.

Experiment 2

Experiment 2 was pre-registered (`https://osf.io/72kqa`).

Method

We used the same language and the same learning and testing procedures in Experiment 2 as in Experiment 1. The two conditions in Experiment 2 were also identical to the ones in Experiment 1. However, Experiment 2 was conducted with a different analysis plan from Experiment 1.

Participants We again 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 18-30 years old, 92 reported being between 30-40, 78 reported being between 40-50, 38 reported being between 50-60, 10 reported being 60 years old or older and two did not report their age. Eligibility criteria, instructions, compensation, and data exclusion protocol were as in Experiment 1. To be eligible for Experiment 2, participants were in addition not allowed to have already taken part in Experiment 1.

Statistical Analysis We built a model to allow for heterogeneity in the participant pool. The model assigns probability r that participants choose markers according to the Wright-Fisher model with selection ("regularizers"). However, the Wright-Fisher model described above is optimal when the selection coefficient is close to zero. To allow for arbitrarily high or low levels of selection, we therefore represent the transition to the output language as a draw from a binomial distribution with parameters n and $f(n,s)$, where n is again population size but $f(n,s)$ is now given by:

$$f(N,s) = \frac{i \cdot e^s}{i \cdot e^s + (n-i)} \quad , \tag{4}$$

where quantities are defined as before.

Our model also assigns probability q that participants choose a single plural marker regardless of input language ("simplifiers"), and probability $1-q-r$ that participants choose plural markers according to a binomial distribution with parameters n and 0.5 ("randomizers"). The best-fit values of q, r, and s for our experimental sample were estimated by selecting the parameter values that maximize the sum of the log-likelihoods of the data.

Note that we did not determine who is a regularizer, simplifier, or randomizer on a subject-by-subject basis. Rather, we determined the most likely frequency of these three types (together with the value of s among regularizers) given all the data. In this way, we were able to simultaneously estimate the selection coefficient among regularizers and the composition of the participant pool. Two-tailed 95% confidence intervals for $\hat{s}$ and $(\hat{r},\hat{q})$ were given by $\ell(s)-\ell(\hat{s}) \leq 1.92$ and $\ell(r,q)-\ell(\hat{r},\hat{q}) \leq 3.98$, respectively.

To test for the effect of drift and selection on noun regularization, we used the following regression model:

$$ln\left(\frac{p}{1-p}\right) = b_0 + b_1 I(f) + b_2 I(s) + b_3 I(f) I(s) \qquad (5)$$

where p is again the proportion of regular nouns, $I(f)$ indicates drift (low: 0; high: 1), $I(s)$ indicates selection (absence: 0; presence: 1), and $I(f)I(s)$ represents the interaction between drift 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).

Results of Experiment 2

Mean completion time in minutes was 8.3 ($s.d. = 5.8$) and 8.4 ($s.d. = 5.6$) for the Drift and Selection Conditions. Data from 10 participants whose completion time was more than two standard deviations from the mean were excluded.

Regularization was higher for low- than for high-frequency nouns in both conditions (Figure 5), with RI estimates for low- and high-frequency 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 ($N = 196$). As expected, selection was negative in the Selection Condition: among regularizers, $\hat{s}$ 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 6). Estimates for low- and high-frequency nouns had roughly the same value.

In the Drift Condition, selection was null among regularizers in the low-frequency class (Figure 6). However, selection among regularizers was negative in the high-frequency class: $\hat{s}$ was $1.97 \pm (0.4, 0.71)$. This could indicate that selection was present in the Drift Condition, but this was likely not the case. In the Drift Condition, the proportion of simplifiers was 0.37 and 0.31, that of randomizers was 0.56 and 0.6, and regularizers made up just 0.07 and 0.09 in low- and high-frequency classes (Figure 7). It is thus most likely that we detected positive selection among regularizers in the high-frequency class simply due to noise, as there were few regularizers in our sample and maximum-likelihood estimation is known to be unreliable in small samples.

In the Selection Condition, on the other hand, the proportion of simplifiers was 0.2 and 0.17, that of randomizers was 0.24 and 0.35, and regularizers made up 0.54 and 0.48 in the low- and high-frequency class. This further corroborates the finding that selection was negative in the Selection Condition and likely absent in the Drift Condition.

Since regularization was higher among low-frequency nouns in both conditions but selection was constant across frequency classes, our results support the hypothesis that low-frequency nouns regularize more because of drift alone. Our regression model provides further support for this finding. Frequency class had a positive effect on noun regularity (0.61 ± 0.21; $p = 0.003$; Table 3). In contrast, selection had a negative effect on noun regularity (-1.05 ± 0.21; $p < 0.0001$). But no interaction between frequency class and selection was detected (0.04 ± 0.3; $p = 0.9$). Our regression model therefore also supports the hypothesis that regularization was driven by drift alone.

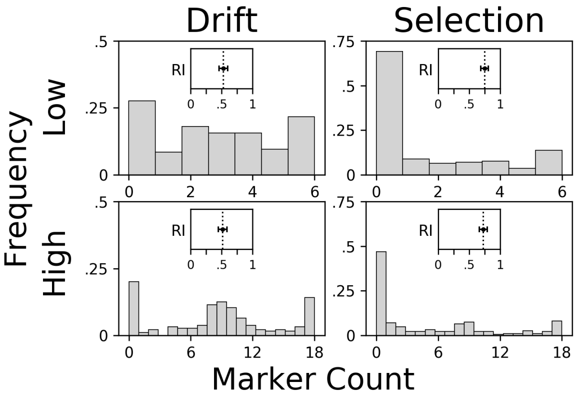

Figure 5: **Marker Counts and Regularization Index (RI) for Experiment 2.** Distribution of irregular marker counts. *Insets:* mean change in proportion of regular nouns between input and output languages (RI) with 95% confidence interval. Drift: $N = 194$. Selection: $N = 196$.

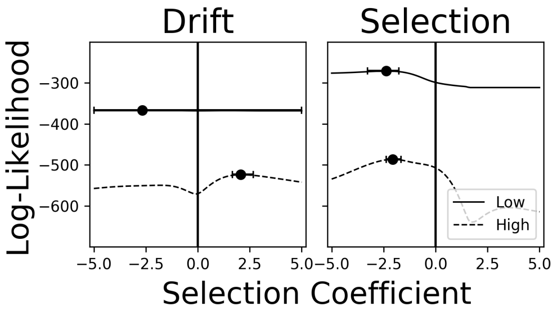

Figure 6: **Sum of log-likelihoods for Experiment 2.** Curves show sum of log-likelihoods given selection coefficient for regularizers in population model; error bars show maximum-likelihood values with 95% confidence intervals, indicating selection against the secondary marker.

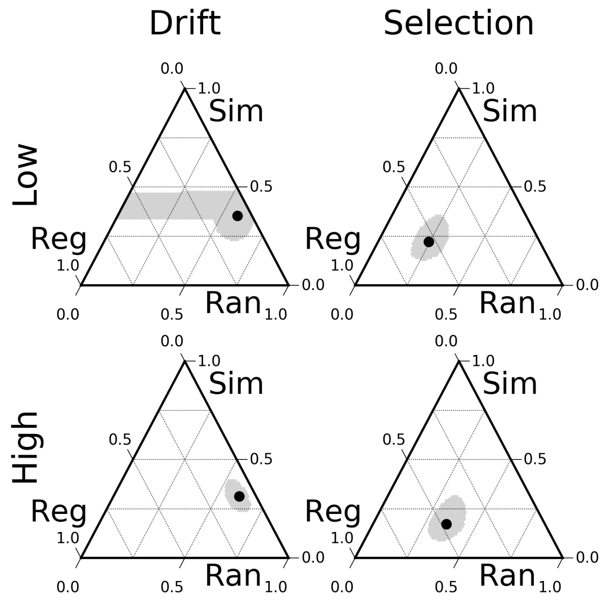

Figure 7: **Population Composition for Experiment 2.** Dots show maximum-likelihood composition in population model with proportion p, q, and $1-p-q$ of randomizers, simplifiers, and regularizers, indicating that the participant pool was heterogeneous and variable across conditions (areas show 95% confidence regions).

Table 3: Logit regression model: $ln(\frac{p}{1-p}) = b_0 + b_1 I(f) + b_2 I(s) + b_3 I(f) I(s)$; see Method for variable definitions. Significant results at the 0.05 level are marked with '*'.

	β	SE	p
intercept (b_0)	0.160343	0.141876	0.2584
frequency (b_1)	0.616502	0.208025	0.003*
selection (b_2)	-1.05572	0.210654	< 0.0001*
freq. × sel. (b_3)	0.0377174	0.296332	0.9

Discussion of Experiment 2

In keeping with Experiment 1, we found that regularization was indeed higher for low- than for high-frequency nouns in both conditions of Experiment 2. We also found that selection for the primary marker was present in the Selection Condition but absent in the Drift Condition, as expected. In the Selection Condition, we further found that the intensity of selection was about the same for both frequency classes. Since selection was constant across frequency classes in the Selection Condition but drift was stronger at low frequencies, our results therefore support the hypothesis that low-frequency nouns underwent more regularization due to drift alone.

Moreover, the participant pool was far from homogeneous with respect to the experimental task. While most participants behaved as expected in regularizing the use of markers in the Selection Condition, many participants did not. Instead, they either randomized their choice of markers or simplified the task by using a single marker. In the Drift Condition, most participants behaved as expected and randomized their choice of markers. But a non-negligible portion of the participants also simplified the task by using a single marker.

General Discussion

Our results show that the difference in regularization between low- and high-frequency nouns observed in Experiments 1 and 2 was due to drift alone. The same might hold for regularization and replacement in natural languages. Our study thus adds to a growing body of evidence suggesting that drift drives this pattern. Our study also highlights the risk of assuming—rather than showing—that participants approach an experimental task as a single homogeneous population.

Some limitations should be noted. First, selection was constant across frequency classes. This means selection could not be responsible for the difference in regularization between frequency classes. In natural languages, however, selection against replacement and regularization may be stronger on high-frequency words if common words function as anchors during language acquisition (cf. Frost, Monaghan, & Christiansen, 2019). Second, factors such as morpheme length or phonological complexity might affect selection strength as well. Even if drift is the primary mechanism of regularization, it may therefore be modulated by selection depending on context. Third, the social context in which language is use is another potential source of selection that was absent in our study. Social meaning and identity, as well as communicative interaction, can also influence which linguistic forms are used and affect the cultural evolution of language (cf. Roberts & Fedzechkina, 2018; Sneller & Roberts, 2018; Galantucci, 2009; Wade & Roberts, 2020). The absence of a social context could therefore explain some of our results, such as the presence of simplifiers. Finally, the artificial language employed in this study was the smallest possible for our purposes. Despite obvious advantages, this also meant that it differed radically from natural languages.

Future work on frequency of use and rates of regularization and replacement might therefore employ more complex languages, incorporate more complex social contexts, and include direct communication between participants (Wade & Roberts, 2020; Sneller & Roberts, 2018) or simulated communication (cf. Buz, Tanenhaus, & Jaeger, 2016). Future work might also implement selection of different strengths across frequency classes or compare different sources of selection (cf. Tamariz, Ellison, Barr, & Fay, 2014). Finally, the integration of natural-language observation, experimental linguistic data, and mathematical models from biology is a strength of our approach, and we are pleased that this has been typical of related research (cf. Karjus, Blythe, Kirby, & Smith, 2020; Newberry et al., 2017). We hope that this kind of interdisciplinary approach be increasingly pursued in the future (cf. Roberts & Sneller, 2020).

Acknowledgements

We are grateful to members of the Akçay-Plotkin lab, the Cultural Evolution of Language Lab, and the Social and Cultural Evolution Working Group for helpful comments on early versions of this study and to Mitchell Newberry for advice on analysis.

References

Buz, E., Tanenhaus, M. K., & Jaeger, T. F. (2016). Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations. *Journal of Memory and Language*, *89*, 68–86.

Frost, R. L. A., Monaghan, P., & Christiansen, M. H. (2019). Mark my words: High frequency marker words impact early stages of language learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(10), 1883.

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

Gray, T. J., Reagan, A. J., Dodds, P. S., & Danforth, C. M. (2018). English verb regularization in books and tweets. *PloS one*, *13*(12), e0209651. (Publisher: Public Library of Science San Francisco, CA USA)

Karjus, A., Blythe, R. A., Kirby, S., & Smith, K. (2018). Challenges in detecting evolutionary forces in language change using diachronic corpora. *arXiv preprint arXiv:1811.01275*.

Karjus, A., Blythe, R. A., Kirby, S., & Smith, K. (2020). Quantifying the dynamics of topical fluctuations in language. *Language Dynamics and Change*, *10*, 86–125.

Karsdorp, F., Manjavacas, E., Fonteyn, L., & Kestemont, M. (2020). Classifying evolutionary forces in language change using neural networks. *Evolutionary Human Sciences*, *2*, 1–40.

Lieberman, E., Michel, J.-B., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature*, *449*(7163), 713.

Newberry, M. G., Ahern, C. A., Clark, R., & Plotkin, J. B. (2017). Detecting evolutionary forces in language change. *Nature*, *551*(7679), 223.

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. (Publisher: Nature Publishing Group)

Reali, F., & Griffiths, T. L. (2010). Words as alleles: connecting language evolution with Bayesian learners to models of genetic drift. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1680), 429–436.

Roberts, G., & Fedzechkina, M. (2018). Social biases modulate the loss of redundant forms in the cultural evolution of language. *Cognition*, *171*, 194–201.

Roberts, G., & Sneller, B. (2020). Empirical foundations for an integrated study of language evolution. *Language Dynamics and Change*, *10*(2), 188–229.

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *116*(3), 444–449.

Sneller, B., & Roberts, G. (2018). Why some behaviors spread while others don't: A laboratory simulation of dialect contact. *Cognition*, *170*, 298–311.

Tamariz, M., Ellison, T. M., Barr, D. J., & Fay, N. (2014). Cultural selection drives the evolution of human communication systems. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1788), 20140488.

Wade, L., & Roberts, G. (2020). Linguistic convergence to observed versus expected behavior in an alien-language map task. *Cognitive Science*, *44*(4), e12829.

Zehr, J., & Schwarz, F. (2018). *PennController for Internet Based Experiments (IBEX)*.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology.* Cambridge: Addison-Wisley.