Supplementary appendix to: Challenges in detecting evolutionary forces in language change using diachronic corpora

Andres Karjus
Centre for Language Evolution, University of Edinburgh, UK
Corresponding author
(a.karjus@sms.ed.ac.uk)

Richard A. Blythe
School of Physics and Astronomy; Centre for Language Evolution, University of Edinburgh, UK

Simon Kirby
Centre for Language Evolution, University of Edinburgh, UK

Kenny Smith
Centre for Language Evolution, University of Edinburgh, UK

This appendix expands on the main text, providing additional information, technical details, and further exploration of the parameter spaces of the models.

A note on corpus annotation quality

While not discussed at length in the main text, the quality of corpus annotation such as lemmatization and part-of-speech tagging plays an equally important role in addition to other corpora-related issues mentioned in the Discussion. Studying the large-scale usage of any linguistic elements of interest relies on the identification of relevant targets in a corpus. Too many erroneously extracted examples can mislead the results. Among the 36 verbs in the sample of Newberry et al, this is especially pertinent for homonymous words like *wet* and *wed*. We already discussed the adjectival usage of *spilt* above. We also found that, for example, 44% of the extracted examples of *wet*.PAST in the first bin (1812-1875 in COHA, under the variable-width binning procedure) were cases of erroneous tagging — being instead other non-past forms of *wet* and occurrences of the adjective *wet*. The same issue applies to *wed*, in addition to being confused with the abbreviation for Wednesday.

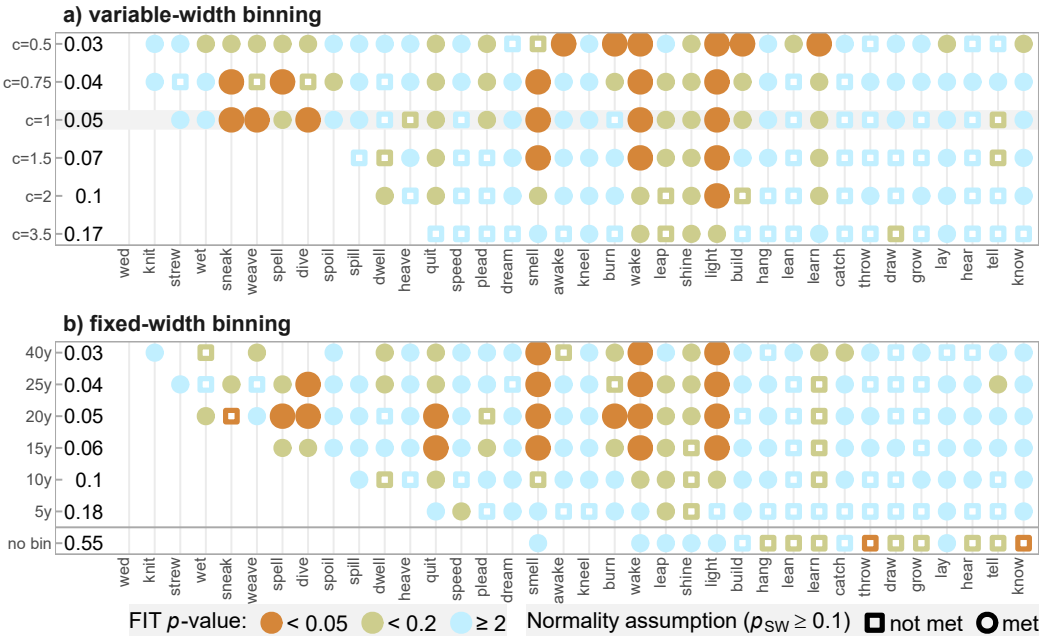

Figure S1: Results of applying the FIT to time series constructed based on 200 years of COHA frequency data. The interpretation of this figure is the same as that of Figure 1, the only difference being the increased minimal within-bin frequency threshold of 100. The constant c determines the number of variable length bins via $n(b) = c\ln(n(v))$. Thus "$c = 1$" corresponds to Newberry et al.'s original results (highlighted with the horizontal grey line). 10y corresponds to fixed bin length of 10 years, etc; 'no bin' refers to no additional binning on top of the default yearly bins in the corpus.

Results based on a different minimal frequency threshold

Figure S1 is intended to complement Figure 1, where we applied a minimal frequency threshold of 10 in each bin (this is mostly relevant for fixed-width binning, as variable-width ensures largely similar bin sizes). Since this is an arbitrary threshold, we also tried a more conservative value of minimal 100 occurrences per bin (for a bin to be included in the time series), with the results reflected in Figure S1. In summary, the higher threshold does not change the results for variable-width binning, besides some lower-frequency verbs being excluded (the empty lower left corner). In fixed-width binning, some results

change, e.g. *spill* is now always flagged as drift, while *burn*, *dive* and *quit* get flagged as selection.

Results of no binning (i.e. using default COHA 1-year bins) should still be taken with a pinch of salt, even when the normality assumption is now met (circles instead of squares) — removing bins with less than 100 tokens leaves even medium-frequency verbs with only a few bins (e.g., 5 in the case of *light*, spread uneven across 200 years; observe also the median bins-to-years ratio of 0.55).

The fact that the minimal threshold affects fixed binning more is not surprising, as the frequencies vary more. This makes variable-width binning a more attractive solution, but its different behaviour should also be taken into consideration. Should the overall frequency of a pair (or set) of variants change over the course of the corpus, it will end up with more bins over the more frequent end of the time scale. As COHA is not uniform in size across time, having considerably less data per year in the first few decades, time series based on variable-width binning of COHA data systematically have longer segments in the beginning and shorter ones towards the end. The "long bins" allow for drawing time series over more sparse corpus segments, where fixed binning would yield unreliably small or empty bins. At the same time, variable-width may by nature gloss over some fluctuations (characteristic of drift) while making a series look more smooth (more characteristic of selection).

Results based on series of different lengths

This figure is intended to complement the simulation section in the main text, which focused on the results of binning a 200-length series into shorter series. Here, no binning is being applied. Figure S2 shows that the FIT produces somewhat different results with the same s given series of different lengths, as expected: when the selection signal is strong enough to be detected (above ~ 0.02), then it is easier to detect it in longer series with more data points than in shorter series. Regardless of series length (at least up to the 200), the false positive rate stays in $[0.03, 0.07]$ (Figure S2.b). This also shows that the higher false positive rate under binning shown in Section 4 does originate in the binning process (which smooths out small fluctuations) rather than simply length difference (binning naturally also making a series shorter).

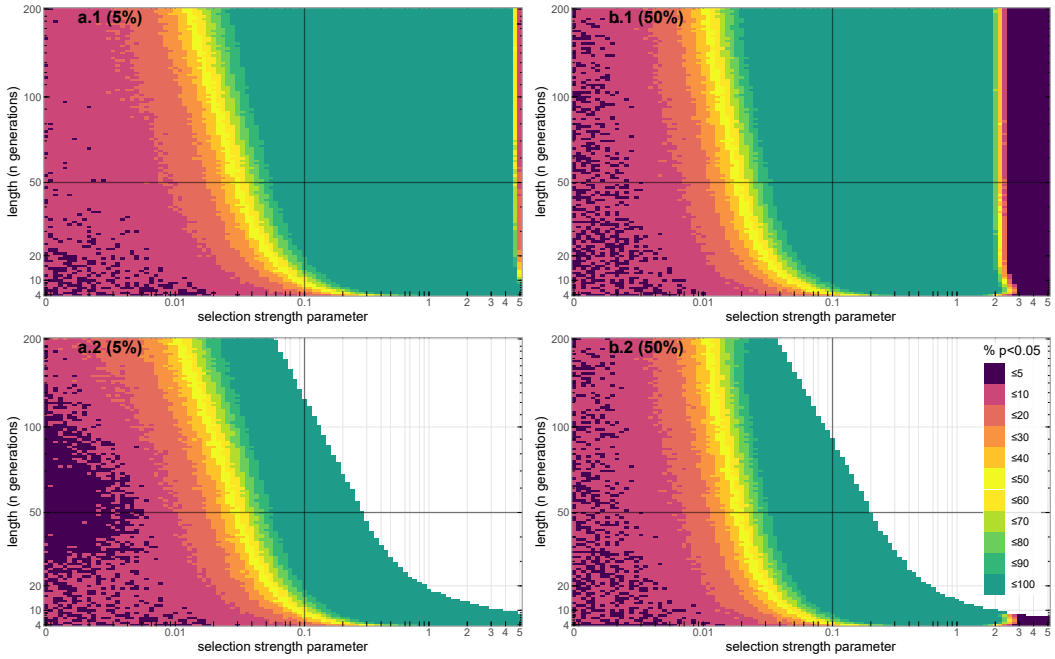

Figure S2: The effect of the interplay of time series length and selection strength s on the results of the FIT. The percentage of FIT $p < 0.05$ (out of 1000 replications for each combination) is reported for a range of time series lengths (y-axis, $[4, 200]$, note the log scale) and the same range of s as above. The left side pair (a) illustrates the case of the time series starting out at 5%, with the 50% condition on the right (b). In the bottom panels (a.2, b.2), series with a Shapiro-Wilk $p < 0.1$ are removed before calculating the percentage. This figure further illustrates the interplay of series length and s that affect the results of FIT.

More examples of the selection coefficient

Figure S3 is intended to complement Figures 3 and 5, where some example Wright-Fisher series where plotted. The s range in our experiments consisted of 200 equidistant values from a log scale between 0.001 and 5, with the addition of 0 in the beginning to be able to explore pure drift.

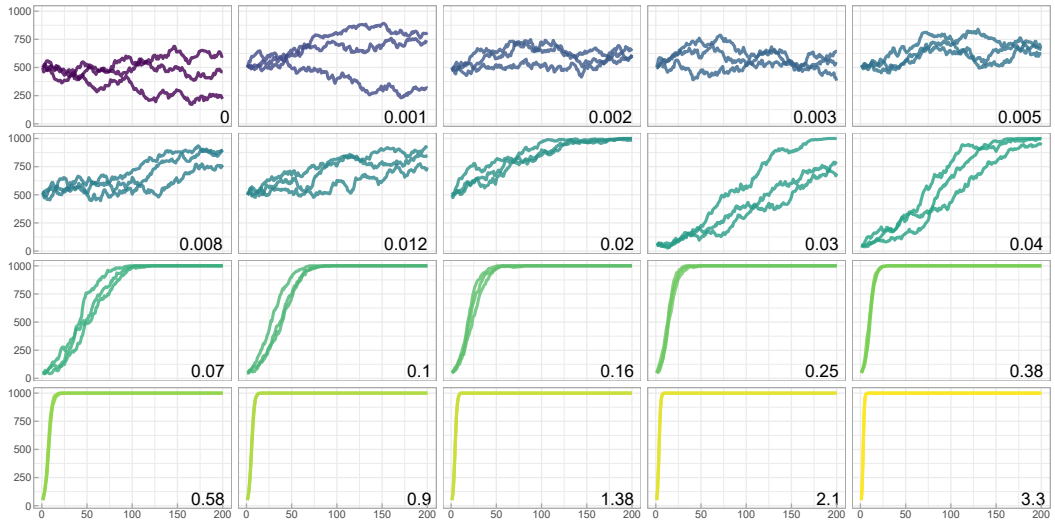

Figure S3: A visualization of the range of selection strength s values explored in the simulation section of this study (shown in the corner of each panel). The horizontal axis corresponds to population size (of the 'mutant' individuals), with time on the horizontal axis. Higher levels of s lead to the mutants taking over the population at faster rates.

The increment normality assumption

The interpretation of the results of the FIT depends how stringently its assumption of the normality of the increments distribution is observed, particularly when s is high. In Figures 4 and 6 we used a Shapiro-Wilk test with a cut-off theshold of 0.1. We conducted additional simulations to see if a lower would yield qualitatively different results, and found it makes very little difference. We also tried using the Lilliefors-Kolmogorov-Smirnov test and the Anderson-Darling test and found all of them to be broadly in agreement: depending on the series starting point and chosen α, the increment normality assumption becomes violated as series (of length 200) approach the s range of 0.05..0.1, with the breaking point being somewhat lower on the s scale in non-binned series and higher in series binned into 10-15 bins (i.e. it is easier to meet the normality assumption if the series is binned).

Increment heteroskedasticity and the Fitness Increment Test

In this additional section, we shed some light on another mathematical aspect of the FIT, the homoskedasticity assumption, as the FIT is, in its core, a one-sample t-test for a zero mean under the assumption of normally-distributed increments with equal variance. For reference, this is the increment transformation process (cf. Section 1.2):

$$Y_i = \frac{v_i - v_{i-1}}{\sqrt{2v_{i-1}(1 - v_{i-1})(t_i - t_{i-1})}} \tag{1}$$

where v_i is the relative frequency of a variant in $(0, 1)$ at time t_i. The rationale behind this rescaling is that, under neutral evolution, the mean increment $v_i - v_{i-1}$ is zero, and its variance is proportional to

$$v_{i-1}(1 - v_{i-1})(t_i - t_{i-1}) \tag{2}$$

However, here we are dealing with estimates of v_{i-1} and v_i obtained from finite samples of size M_{i-1} and M_i, respectively. This leads to additional contributions to the variance of the increment $v_i - v_{i-1}$, arising from the variance of the binomial distribution $v(1 - v)/M$, where v is the mean value and M is the sample size. To a first approximation, the total variance of the increment is obtained by summing the three contributions. That is, $(v_i - v_{i-1})$ has a variance of

$$\frac{v_{i-1}(1 - v_{i-1})(t_i - t_{i-1})}{N} + \frac{v_{i-1}(1 - v_{i-1})}{M_{i-1}} + \frac{v_i(1 - v_i)}{M_i} \, . \tag{3}$$

where N stands for effective population size. The transform divides all of this by $v_{i-1}(1 - v_{i-1})(t_i - t_{i-1})$ which leads to a variance of each *transformed* increment Y_i of approximately

$$\frac{1}{N} + \frac{1}{M_{i-1}(t_i - t_{i-1})} + \frac{v_i(1 - v_{i-1})}{M_i(t_i - t_{i-1})v_{i-1}(1 - v_{i-1})} \, . \tag{4}$$

The FIT can be expected to perform as intended when this variance is constant. This is the case when $1/M \ll (t_i - t_{i-1})/N$ or $M(t_i - t_{i-1}) \gg N$ (assuming N can be inferred, which is not trivial, but cf. Newberry et al.). The transformed increments based on corpus data basically never have perfectly equal variance once sample size is taken into account, but will be roughly constant when the sample sizes are large (relative to N). The variable-width

binning, as employed by Newberry et al., assures that the variances are more or less equal, as each bin has roughly the same number of tokens. The worry with fixed-width binning — including the default data binning of one year in COHA as well further binning of the years into decades and so on — is that the variance is not going to be equal, as bins may or may not cover a similar number of tokens.

We calculated these values for the English verb data (with the simplification of excluding N, which is not trivial to infer). Variable-width binning consistently yields small increment variances with a very small standard deviation (depending in turn on the variance in the bin sizes in the original data). Using the data without further binning (i.e. 1-year bins from COHA) yields multiple magnitudes higher values for both, as does fixed binning into short bins. But starting at decade-length bins (for higher-frequency verbs like *light*) and 20-year bins (for lower-frequency verbs like *spell*), as bin sizes approach 100 tokens, the picture becomes quite similar to variable-width binning.

It is not clear, however, how much heteroskedasticity is bad enough to lead to spurious results. For example, is it invalid to interpret the results of the FIT based on 1-year or 5-year bins at all, given typical sample sizes in a corpus like COHA? While this would benefit from more through future investigation, we attempt to shed some light on this by conducting more Wright-Fisher simulations where we manipulate the size of M in each generation, and the standard deviation of sample sizes (σ_S), as well as apply different binning strategies to the resulting time series. In Figure S4, the series length is 200 as in the previous simulations, $s = 0$ (as we are interested in the false positives rate), and we explore two N sizes, 10000 (left side column in Figure S4) and 1000 (right side). Each pixel represents the share of FIT $p < 0.05$ in 1000 replications with the given parameter combination.

For each replication in a combination, we run a Wright-Fisher simulation, but to construct the time series, take a random sample of individuals M at each of the 200 generations. The sample sizes are in turn generated by sampling values from a log-normal distribution with a mean of $\ln(M)$ (y-axis in Figure S4) and $\sigma_S \in \ln([1, 2])$ (corresponding to the x-axis in Figure S4). The log-normal distribution excludes 0, but with a high standard deviation, some of the generated M values can exceed that of the population size, so when reconstructing the time series, they are truncated by taking $\min(M, N)$. After this manipulation however, where M and V are both high, the resulting actual standard deviation across the 200 M sample sizes would not correspond to the predetermined parameter of σ_S, therefore, such replications are filtered out (along with series where the normality assumption is violated, at Shapiro-Wilk

$p < 0.1$). When less than 10% of the replications for a combination are valid, it is excluded from plotting (the white areas in Figure S4).

The leftmost column of pixels on each panel corresponds to no variance in sample sizes, i.e., the samples are of equal size, corresponding exactly to the value on the y-axis. The top left pixel therefore represents the baseline Wright-Fisher simulation result with no downsampling (the false positive rate being around 5% in both N at $\alpha = 0.05$). Each top panel shows the results without binning, with the lower ones showing results when the series are binned into a smaller number of bins (after the aforementioned downsampling procedure).

In Figure S4, where cold blueish colours represent percentages of FIT $p < 0.05$. Ideally, as $s = 0$, all of the panels should be devoid of any warm colours. Looking at any single panel, the columns of pixels right of the no-variance leftmost column are not any more yellow than the leftmost column. This demonstrates that variance in sample sizes does not make any discernable difference — it does not make the already borderline false positive rate any worse. This observation holds between binning choices. Binning itself does increase the false positive rate, as already determined in Section 4 (panels below the top ones exhibit more yellow). If anything, it would seem series based on samples of size $M < N$ and increased σ_S have an improved (i.e. smaller) false positive rate, an effect particularly pronounced when the series are binned. This is however an expected result stemming from the added sampling noise (making any series look more "random" to the test).

The heteroskedasticity question remains somewhat unresolved, but based on these results we can say that at least the false positive rate of FIT is unlikely to be considerably affected by differing bin sizes. In terms practical guidelines, to be safe, if applicable variable-width binning should be used with FIT as proposed by Newberry et al. If fixed-width binning is used, then bins should consist of 100 occurrences or more. In the end this is not only a variance problem, but a small sample size problem. A large number of bins consisting each of only tens of occurrences has considerable sampling noise. Given the same corpus, a small number of bins consisting each of hundreds of occurrences can gloss over the true trajectory of change, but also any statistical test based on too few data points is unreliable. In other words, there's no data like more data.

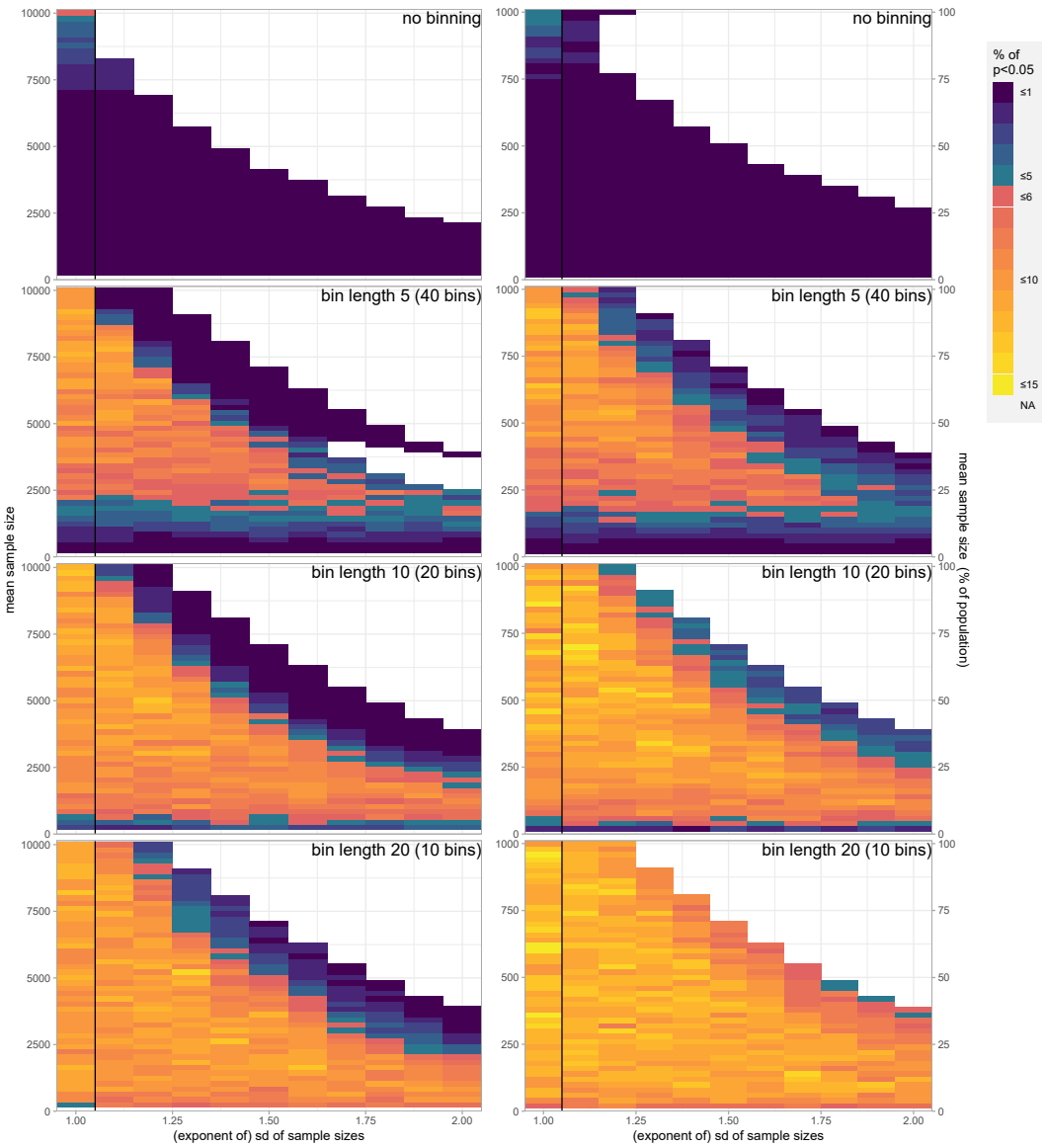

Figure S3: False positive rates of the FIT based on Wright-Fisher simulations with downsampled populations. Column of panels on the left: $N = 10000$. Panels on the right: $N = 1000$. The cool colours correspond to percentages of $p < 0.05$ below 5%, warm colours indicate higher percentages. This figure illustrates that while binning tends to introduce more false positives, in any given binning strategy, added variance in the underlying occurrence counts (and thus bin sizes) does not.