doi:10.1038/nature24455

Detecting evolutionary forces in language change

Mitchell G. Newberry[1]*, Christopher A. Ahern[2]*, Robin Clark[2] & Joshua B. Plotkin[1]

Both language and genes evolve by transmission over generations with opportunity for differential replication of forms[1]. The understanding that gene frequencies change at random by genetic drift, even in the absence of natural selection, was a seminal advance in evolutionary biology[2]. Stochastic drift must also occur in language as a result of randomness in how linguistic forms are copied between speakers[3,4]. Here we quantify the strength of selection relative to stochastic drift in language evolution. We use time series derived from large corpora of annotated texts dating from the 12th to 21st centuries to analyse three well-known grammatical changes in English: the regularization of past-tense verbs[5–9], the introduction of the periphrastic 'do'[10], and variation in verbal negation[11]. We reject stochastic drift in favour of selection in some cases but not in others. In particular, we infer selection towards the irregular forms of some past-tense verbs, which is likely driven by changing frequencies of rhyming patterns over time. We show that stochastic drift is stronger for rare words, which may explain why rare forms are more prone to replacement than common ones[6,9,12]. This work provides a method for testing selective theories of language change against a null model and reveals an underappreciated role for stochasticity in language evolution.

There is a rich history of exchange between linguistics and evolutionary biology[13–15]. Linguists have uncovered notable regularities in language change by examining which new forms enter a language and which forms are lost[9,11,13,16]. Massive digital corpora[7,17] now provide precise frequency time series of one form replacing another, which enable us to quantify evolutionary forces in language change using methods drawn from population genetics.

Language change involves competition between alternative linguistic forms (such as 'sneaked' versus 'snuck') that may differ according to sound, morphology, or syntactic structure[1,6,12,18–20]. With every utterance, a speaker either invents a new form or uses one copied from other speakers. Forces that bias a speaker towards adopting one form instead of another have been documented in detail[21]; examples include phonological analogy[9,22], over-emphasis[11,23], and a host of other social and cognitive factors[18,24]. Any such bias in copying constitutes a form of selection in language evolution[14]. Explanations for language change, in which one linguistic form increases in frequency and eventually replaces an alternative form over the course of generations, uniformly invoke selective mechanisms[19,25]. However, the frequencies of alternative variants can change markedly even without bias towards one form or another, as a result of stochastic drift: randomness in the set of forms that each speaker happens to encounter and reproduce (Fig. 1). To determine the importance of directional forces we must first assess whether an observed language change is consistent with stochasticity in propagation alone.

Drift is recognized as an important null hypothesis in population genetics[26] and cultural evolution[1,27]. More recently, linguists have suggested the use of null models for language change. Several models, including neutral evolution[28,29], have been proposed[3] and some changes (such as new dialect formation) have been attributed to stochastic drift[30]. However, methods to analyse drift versus selection in available linguistic data have not yet been developed.

Here we systematically quantify the contributions of drift and selection to three well-known grammatical changes in English: the development of the morphological past tense in contemporary American English[5,31] (spilt → spilled); the rise of the periphrastic 'do' in Early Modern English[10] ('You say not' → 'You do not say'); and Jespersen's cycle of sentential negation in Middle English[11] ('Ic ne secge' → 'I ne seye not' → 'I say not'). Our analyses draw on annotated texts that range in time from the Norman conquest of England to the 21st century. In each case, we test whether observed linguistic changes are consistent with stochastic drift or must involve directional forces. We compare the frequencies of alternative linguistic variants over time to predictions under the Wright–Fisher model of neutral stochastic drift. This model was first introduced in population genetics[2] but it has also been derived as a null model of linguistic change under Bayesian learning[4], in which the inverse of the population size parameter N governs the amount of stochasticity in transmission and thus the strength of drift.

We analysed the evolution of past-tense verb conjugation by collecting verb tokens from the Corpus of Historical American English[17]. This corpus comprises more than four hundred million words, tagged for part of speech, from over one hundred thousand texts dated between the years 1810 and 2009. From all tokens tagged as having the simple past tense, we selected those lemmas with two past-tense variants that each occurred at least 50 times in the corpus (Supplementary Information section 1.4). This produced 704,081 tokens, which provide frequency time series for the regular versus irregular forms of 36 polymorphic verbs (Fig. 2). These verbs range from the very rare ('wed', one in two million) to common ('know', one in two thousand). For each time series we computed a two-sided P value for rejecting neutral stochastic drift by the frequency increment test[32] (FIT, see Supplementary Information section 1.2). We also inferred the most likely population size (N) and selection coefficient (s) in favour of one linguistic variant over another (Extended Data Table 1 and Supplementary Information section 1.3).

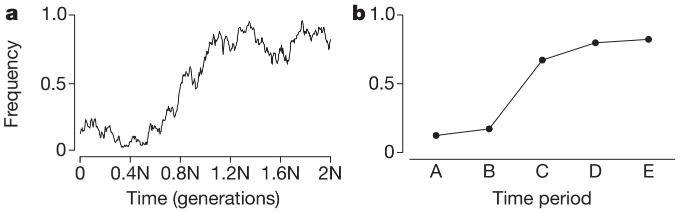


Figure 1 | A null model of language change. Stochastic drift, random fluctuations in the frequencies of alternative forms, can accumulate to produce substantial change over time. **a**, Example time series of frequency variation produced by the neutral Wright–Fisher model of stochastic drift with population size parameter N. Although the complete time series shows random fluctuations, linguistic time series are typically binned into time periods. **b**, Binning the time series in **a** produces a characteristic S-shaped curve, which is often accepted as evidence of a directional force favouring one linguistic variant over others[19,25]. This example illustrates the need to test hypotheses against a null model to definitively infer the presence of selective forces in language change[29].

[1]Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [2]Department of Linguistics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.
*These authors contributed equally to this work.

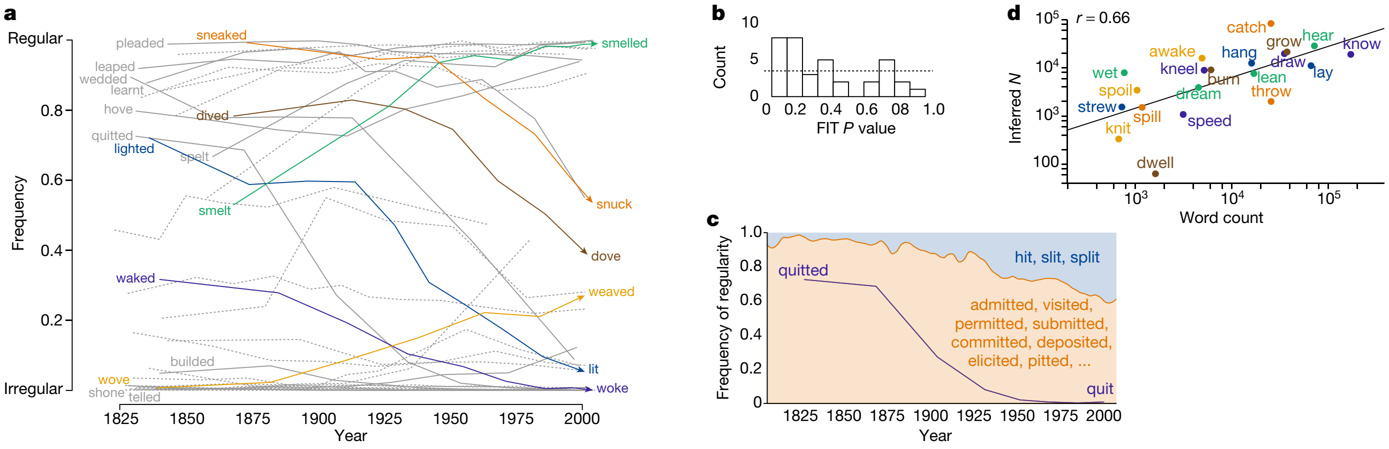


Figure 2 | Verb regularization and irregularization. We analysed 36 verbs with multiple past-tense forms that appear in the Corpus of Historical American English[17]. **a**, Six of these verbs (coloured lines) experience selection for either regularization or irregularization, each with nominal $P < 0.05$ by the frequency increment test (FIT) of selection (false-discovery rate = 30%, Extended Data Table 1). The regular form is favoured in two of these cases and the irregular form in the remaining four cases. Ten more verbs (solid grey lines), of which four are regularizing, are significant at specificity $1 - \alpha = 0.8$ with false-discovery rate = 45%. **b**, The distribution of nominal FIT P values is non-uniform (Kolmogorov–Smirnov $P = 0.002$), which confirms that some verbs experience selection. **c**, Changing use of rhyming patterns may drive selection for irregular forms, such as quitted → quit, for which irregularization coincides with the increasing use of the irregular verbs 'hit', 'slit', and 'split' (Extended Data Fig. 1). **d**, Among the remaining 20 verbs most consistent with neutrality (**a**, dashed grey lines), the log inferred population size (assuming $s = 0$) correlates with log token count in the corpus (Pearson's $r = 0.66$, $P = 0.002$).

We conclude that selection is driving changes in past-tense conjugation for six of the polymorphic verbs, each with nominal $P < 0.05$. In four of these cases selection favours the irregular variant (lighted → lit, waked → woke, sneaked → snuck, dived → dove); the two remaining cases exhibit regularization (wove → weaved, smelt → smelled). The false-discovery rate among these six verbs is 30% (Fig. 2b). Furthermore, we note that the distribution of all FIT P values is non-uniform (Kolmogorov–Smirnov $P = 0.002$, Fig. 2b), which confirms that selection is operating on some of the polymorphic verbs.

Selection for regularization comes as no surprise; prevailing linguistic theory predicts regularization[5,9] for reasons of economy or cognitive ease[5,33,34]. Trends towards past-tense regularization have previously been observed, especially for rare words over long timescales from Old to Modern English[6,7,9]. We identify cases of incipient regularization (such as wove → weaved), in which the regular variant is in the minority at present but is predicted by our analysis to eventually replace the irregular form.

Selection for irregularization is more mysterious, although several cases have been noted[7,16,31]. In Modern English, we find that irregularization is as common as regularization (Fig. 2). One possible explanation involves rhyming. Psychological studies have found speakers willing to copy or invent irregular variants (such as spling/splung[22]) that rhyme with existing irregular verbs[35]. Our analysis of 'dived' versus 'dove' as the past-tense conjugation of 'dive' reveals selection for dive/dove, which coincides with a marked increase in the use of the irregular verb drive/drove in the corpus, associated with the invention of cars in the 20th century. More generally, in all eleven cases (light, dive, quit, tell, leap, build, kneel, know, throw, knit and grow; see Extended Data Fig. 1, Extended Data Table 2, and Supplementary Information section 1.5) the inferred selection coefficient (s) favours the irregular variant of a polymorphic past-tense verb if similar-sounding irregular verbs are on the rise in the corpus. For example, selection for quitted → quit coincides with the increased use of the irregular verbs hit/hit and split/split (Fig. 2c). The frequency of 'split' increased nearly fourfold over the past century, as split acquired an additional meaning (to leave or depart). Thus, a semantic change in one irregular verb (split) may have induced selection for irregularization in another, semantically unrelated verb (quit) that shares the same present/past rhyming pattern.

Selection towards an irregular variant can also occur when similar-sounding irregular verbs are on the decline, as in the case of wedded → wed (Extended Data Fig. 1). Our inference of selection for wedded → wed is notable because it contradicts previous work that predicted the regularization of wed → wedded on the basis of long-term trends[6].

Drift alone is sufficient, however, to explain the observed changes for the majority of verbs we analysed in Modern English (Extended Data Table 1, FIT P values). These include verbs previously described as undergoing regularization, such as spilt → spilled and burnt → burned[7,31]. Failure to reject neutrality in these cases does not imply that selection is entirely absent. For example, there is probably some selection for knitted → knit due to rhyming, as with quitted → quit (Extended Data Fig. 1). Nonetheless, the inferred strength of selection for 'knit' is too weak relative to drift to affect its dynamics: $|Ns| = 1.67$ for 'knit' (FIT $P = 0.76$) in contrast to $|Ns| = 30.51$ for 'lit' (FIT $P = 0.003$) (Extended Data Table 1). Even with some amount of selection, if drift is strong enough the dynamics are indistinguishable from neutral[2].

Among the verbs with dynamics dominated by drift, the strength of drift correlates inversely with the overall frequency of the verb in the corpus (Fig. 2). This result implies that common words should exhibit less variability over time than rare words, a phenomenon that has been observed in a number of empirical studies[6,9,12] and previously attributed to stronger purifying selection against novel variants of common words[12]. Our analysis provides an alternative and complementary explanation for faster rates of replacement in rare words: whether under selection or not, rare words experience more stochasticity in transmission. Our explanation further predicts that for rare words the replacement of one form by another is more likely to occur by random chance, whereas such substitutions in common words are more likely to be caused by selection.

Next we analysed the rise of *do*-support in Early Modern English[10], as the auxiliary verb 'do' came to express the tense of a sentence. Over the course of centuries, for example, 'You say not' became 'You do not say' and 'Say you?' became 'Do you say?'. We collected instances of potential *do*-support from the Penn Parsed Corpora of Historical English (Supplementary Information section 1.6). This dataset includes roughly seven million syntactically parsed words from 1,220 texts of British English, and it offers a much larger time series than those used in previous work[10]. We extracted 20,729 instances of potential *do*-support in the context of affirmative questions, negative questions, negative declaratives, and negative imperatives.

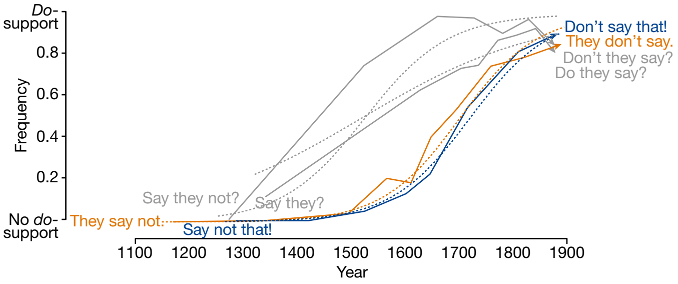


Figure 3 | The rise of the periphrastic 'do' in Early Modern English. The frequency of 'do' as an auxiliary verb first rose in the context of interrogative sentences (grey). However, we cannot reject drift for either affirmative interrogatives (FIT $P=0.18$) or negative interrogatives (FIT $P=0.27$). Subsequently, *do*-support rose rapidly in negative declarative and negative imperative sentences, where we detect selection (FIT $P=0.005$ and $P=0.003$, respectively). Dotted lines plot the logistic curve with slope determined by the maximum-likelihood selection coefficient inferred in each grammatical context (Extended Data Table 3). These results suggest *do*-support rose by chance through drift in interrogative statements, setting the stage for directional evolution of *do*-support in other grammatical contexts.

We find that the rise of the periphrastic 'do' was more rapid in negative declarative and imperative statements, for which we reject drift (FIT $P=0.005$ and $P=0.003$, respectively), than it was in affirmative questions, for which we fail to reject drift (FIT $P=0.18$, Fig. 3). *Do*-support also appears to rise rapidly in negative questions, although in this case the force of drift is strongest (Extended Data Table 3) and so we fail to reject drift (FIT $P=0.27$, Fig. 3) despite sufficient power (Supplementary Information section 1.6). We might expect that selection for an auxiliary verb would operate equally in all grammatical contexts[19], and yet the extensive parsed corpora do not support this hypothesis. Our analysis suggests an alternative scenario: the periphrastic 'do' first drifted by chance to high frequency in questions, which then induced a directional bias towards 'do' in declarative and imperative statements for reasons of grammatical consistency or cognitive ease.

Finally, we studied the evolution of syntactic verbal negation from the 12th to the 16th centuries, using 5,475 negative declaratives extracted from the Penn–Helsinki Parsed Corpus of Middle English. We observe pre-verbal negation (for example, Old English 'Ic ne secge') giving way to embracing bipartite negation (Middle English 'I ne seye not') and then finally to post-verbal negation (Early Modern English 'I say not'), in a pattern known as Jespersen's cycle[11]. For both transitions that form this cycle, we reject neutral drift (FIT $P<0.05$, Fig. 4). This provides statistical support for longstanding hypotheses that changes in verbal negation are driven by directional forces, such as phonetic weakening[11], or a tendency for speakers to over-use more emphatic forms of negation[11,23] that then lose emphasis as they become dominant[23,36]. Although directionality in Jespersen's cycle was first recognized by comparing multiple languages[11], we reach the same conclusion by analysing changes in English alone.

Methods drawn from phylogenetics have enabled researchers to infer the relationships among divergent languages[12,37–40]. By contrast, the study of how a language changes over short timescales has not taken full advantage of statistical inference. Yet changes within a language must be the origin of differentiation between languages[41]. Combining massive digital corpora with time series inference techniques from population genetics now allows us to disentangle distinct forces that drive language evolution. How exactly individual-level cognitive processes in a language learner[5,11,19,33,34] produce population-level phenomena, such as drift and selection[42], remains a topic for future research.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Code Availability Source code is available online at http://github.com/mnewberry/ldrift.

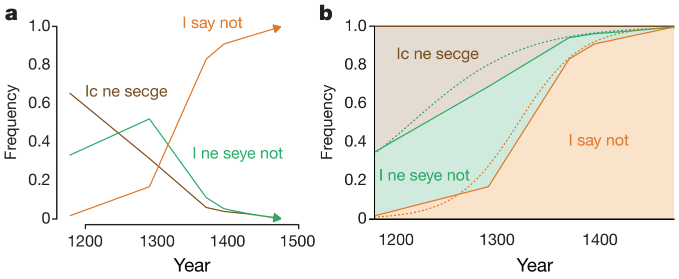


Figure 4 | Evolution of verbal negation. Pre-verbal negation (for example, Old English 'Ic ne secge') gave way to embracing bipartite negation (Middle English 'I ne seye not') and then to post-verbal negation (Early Modern English 'I say not'). **a**, Frequencies of pre-verbal, bipartite, and post-verbal forms among 5,918 instances of negation from 56 texts in the Penn–Helsinki Parsed Corpus of Middle English. **b**, We infer selection for bipartite and post-verbal negation in the background of pre-verbal forms (FIT $P=0.02$) and selection for post-verbal negation in a mixed population of pre-verbal and bipartite forms (FIT $P=0.04$). Dotted lines indicate logistic curves corresponding to maximum-likelihood selection coefficients.

Data Availability Data that support the findings of this study are available as the Corpus of Historical American English (https://corpus.byu.edu/coha/) and the Penn Parsed Corpora of Historical English (http://www.ling.upenn.edu/hist-corpora/).

Received 28 February; accepted 22 September 2017.

Published online 1 November 2017.

1. Cavalli-Sforza, L. & Feldman, M. *Cultural Transmission and Evolution: A Quantitative Approach* (Princeton Univ. Press, 1980).
2. Crow, J. F. & Kimura, M. *An Introduction to Population Genetics Theory* (Harper & Row, 1970).
3. Bentley, R. A., Hahn, M. W. & Shennan, S. J. Random drift and culture change. *Proc. R. Soc. Lond. B* **271,** 1443–1450 (2004).
4. Reali, F. & Griffiths, T. L. Words as alleles: connecting language evolution with Bayesian learners to models of genetic drift. *Proc. R. Soc. Lond. B* **277,** 429–436 (2010).
5. Pinker, S. Rules of language. *Science* **253,** 530–535 (1991).
6. Lieberman, E., Michel, J.-B., Jackson, J., Tang, T. & Nowak, M. A. Quantifying the evolutionary dynamics of language. *Nature* **449,** 713–716 (2007).
7. Michel, J.-B. *et al.* Quantitative analysis of culture using millions of digitized books. *Science* **331,** 176–182 (2011).
8. Reali, F. & Griffiths, T. L. The evolution of frequency distributions: relating regularization to inductive biases through iterated learning. *Cognition* **111,** 317–328 (2009).
9. Hooper, J. B. in *Current Progress in Historical Linguistics: Proc. 2nd International Conference on Historical Linguistics* (ed. Christie, W. M.) 96–105 (North-Holland, 1976).
10. Ellegård, A. *The Auxiliary Do: The Establishment and Regulation of its Use in English* (Almqvist & Wiksell, 1953).
11. Jespersen, O. *Negation in English and Other Languages* (AF Høst, 1917).
12. Pagel, M., Atkinson, Q. D. & Meade, A. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* **449,** 717–720 (2007).
13. Schleicher, A. *Darwinism Tested by the Science of Language* (John Camden Hotten, 1869).
14. Darwin, C. *The Descent of Man, and Selection in Relation to Sex* (Murray, 1888).
15. Haeckel, E. *Natürliche Schöpfungs-Geschichte* (Georg Reimer, 1868).
16. Bybee, J. L. & Moder, C. L. Morphological classes as natural categories. *Language* **59,** 251–270 (1983).
17. Davies, M. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora* **7,** 121–157 (2012).
18. Labov, W. *Principles of Linguistic Change* Vol. 3 (Blackwell, 2010).
19. Kroch, A. S. Reflexes of grammar in patterns of language change. *Lang. Var. Change* **1,** 199–244 (1989).
20. Christiansen, M. H., Chater, N. & Culicover, P. W. *Creating Language: Integrating Evolution, Acquisition, and Processing* (MIT Press, 2016).
21. Croft, W. *Explaining Language Change: An Evolutionary Approach* (Pearson Education, 2000).
22. Prasada, S. & Pinker, S. Generalisation of regular and irregular morphological patterns. *Lang. Cogn. Process.* **8,** 1–56 (1993).
23. Dahl, O. *Inflationary Effects in Language and Elsewhere* (John Benjamins, 2001).

24. Hawkins, J. A. A parsing theory of word order universals. *Linguist. Inq.* **21,** 223–261 (1990).
25. Blythe, R. A. & Croft, W. S-curves and the mechanisms of propagation in language change. *Language* **88,** 269–304 (2012).
26. Wright, S. Evolution in Mendelian populations. *Genetics* **16,** 97–159 (1931).
27. Kandler, A. & Shennan, S. A non-equilibrium neutral model for analysing cultural change. *J. Theor. Biol.* **330,** 18–25 (2013).
28. Hahn, M. W. & Bentley, R. A. Drift as a mechanism for cultural change: an example from baby names. *Proc. R. Soc. Lond. B* **270,** S120–S123 (2003).
29. Blythe, R. A. Neutral evolution: a null model for language dynamics. *Adv. Complex Syst.* **15,** 1150015 (2012).
30. Baxter, G. J., Blythe, R. A., Croft, W. & McKane, A. J. Modeling language change: an evalution of Trudgill's theory of the emergence of New Zealand English. *Lang. Var. Change* **21,** 257–296 (2009).
31. Cuskley, C. F. *et al.* Internal and external dynamics in language: evidence from verb regularity in a historical corpus of English. *PLoS One* **9,** e102882 (2014).
32. Feder, A. F., Kryazhimskiy, S. & Plotkin, J. B. Identifying signatures of selection in genetic time series. *Genetics* **196,** 509–522 (2014).
33. Jakobson, R., Waugh, L. R. & Monville-Burston, M. *On Language* (Harvard Univ. Press, 1995).
34. Zipf, G. K. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology* (Addison–Wesley, 1949).
35. Ullman, M. T. Acceptability ratings of regular and irregular past-tense forms: evidence for a dual-system model of language from word frequency and phonological neighbourhood effects. *Lang. Cogn. Process.* **14,** 47–67 (1999).
36. Crawford, V. P. & Sobel, J. Strategic information transmission. *Econometrica* **50,** 1431–1451 (1982).
37. Ringe, D., Warnow, T. & Taylor, A. Indo-European and computational cladistics. *Trans. Philol. Soc.* **100,** 59–129 (2002).
38. Gray, R. D. & Atkinson, Q. D. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426,** 435–439 (2003).
39. Pagel, M. in *The Princeton Guide to Evolution* (ed. Losos, J.) Ch. VIII.9 (Princeton Univ. Press, 2013).
40. Pagel, M. Human language as a culturally transmitted replicator. *Nat. Rev. Genet.* **10,** 405–415 (2009).
41. Lupyan, G. & Dale, R. in *Language Structure and Environment: Social, Cultural, and Natural Factors* (eds De Busser, R. & LaPolla, R. J.) Ch. 11 (2015).
42. Tamariz, M., Ellison, T. M., Barr, D. J. & Fay, N. Cultural selection drives the evolution of human communication systems. *Proc. R. Soc. Lond. B* **281,** 20140488 (2014).

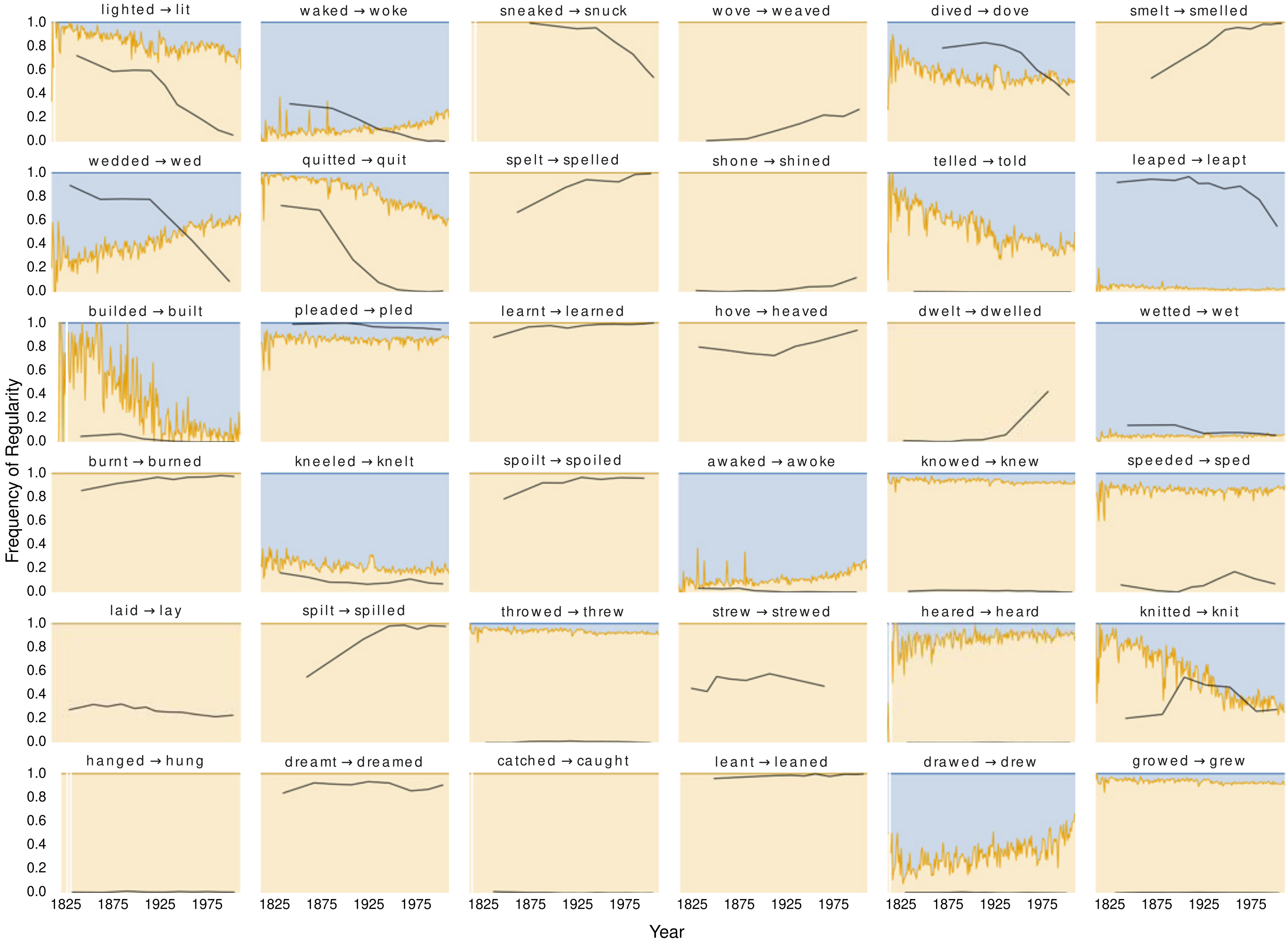
Supplementary Information is available in the online version of the paper.

Acknowledgements We thank H. Bacovcin, T. Kroch, and M. Liberman. R.C. acknowledges support from the University of Pennsylvania Research Foundation. J.B.P. acknowledges support from the David & Lucile Packard Foundation, the US Defense Advanced Research Projects Agency (D12AP00025), and the US Army Research Office (W911NF-12-1-0552).

Author Contributions M.G.N., C.A.A., R.C., and J.B.P. conceived the study, designed the analysis, and wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to J.B.P. (jplotkin@sas.upenn.edu).

Reviewer Information *Nature* thanks R. A. Bentley and the other anonymous reviewers for their contribution to the peer review of this work.



Extended Data Figure 1 | Time series of changing rhyming patterns. Each panel shows the time series of a polymorphic verb (black lines), repeated from Fig. 2a, and the frequency of similar-sounding monomorphic regular (orange) and irregular (blue) verbs in the Corpus of Historical American English. The tokens included are all tenses of those lemmas that possess a pronunciation known to the Carnegie Mellon University Pronouncing Dictionary in both the lemma and the simple past tense. The list of verbs incorporated in each time series is given in Extended Data Table 2. For 17 polymorphic verbs we find no similar-sounding monomorphic irregular verbs (all-orange panels). The title of each panel indicates the sign of the maximum-likelihood selection coefficient, either regular → irregular or irregular → regular.

Extended Data Table 1 | FIT results for past-tense verbs

Lemma	Regular	Irregular	Count	FIT p-value	Inferred N	Inferred s	Inferred $N_{s=0}$
light	lighted	lit	8,869	0.0030	1,247	-0.024	770
wake	waked	woke	7,186	0.0055	1,714	-0.024	4,005
sneak	sneaked	snuck	898	0.0150	4,135	-0.039	29
weave	weaved	wove	907	0.0211	1,052	0.013	126
dive	dived	dove	1,036	0.0477	2,148	-0.018	710
smell	smelled	smelt	4,555	0.0495	974	0.038	1,708
wed	wedded	wed	211	0.08	888	-0.026	154
quit	quitted	quit	2,734	0.10	346	-0.048	430
spell	spelled	spelt	962	0.10	1,556	0.025	2,055
shine	shined	shone	8,424	0.10	20,110	0.019	989
tell	telled	told	129,041	0.12	20,930	-0.050	314,100
leap	leaped	leapt	8,336	0.13	2,003	-0.019	346
build	builded	built	9,109	0.14	87	-0.063	8,602
plead	pleaded	pled	3,810	0.14	8,050	-0.006	3,756
learn	learned	learnt	18,851	0.16	2,467	0.027	6,918
heave	heaved	hove	2,392	0.20	2,371	0.008	3,663
dwell	dwelled	dwelt	1,621	0.20	16,170	0.033	63
wet	wetted	wet	770	0.24	5,707	-0.009	7,903
burn	burned	burnt	6,097	0.24	7,213	0.014	9,045
kneel	kneeled	knelt	5,185	0.30	7,299	-0.006	8,912
spoil	spoiled	spoilt	1,045	0.30	3,519	0.018	3,431
awake	awaked	awoke	4,926	0.32	1,676	-0.036	15,860
know	knowed	knew	171,518	0.34	9,397	-0.007	18,980
speed	speeded	sped	3,142	0.39	867	-0.002	1,077
lay	laid	lay	66,436	0.45	10,610	-0.002	11,070
spill	spilled	spilt	1,178	0.47	1,266	0.031	1,509
throw	throwed	threw	25,612	0.61	729	-0.011	2,001
strew	strewed	strew	727	0.66	1,539	0.000	1,537
hear	heared	heard	72,052	0.72	1,129	-0.033	28,530
knit	knitted	knit	675	0.76	173	-0.010	333
hang	hanged	hung	16,079	0.77	3,855	-0.012	12,450
dream	dreamed	dreamt	4,530	0.77	2,832	0.005	3,907
catch	catched	caught	25,529	0.78	22,520	-0.021	83,060
lean	leaned	leant	16,981	0.85	2,166	0.013	7,594
draw	drawed	drew	35,213	0.87	2,102	-0.026	19,620
grow	growed	grew	37,444	0.93	7,094	-0.013	21,340

We analysed 36 verbs with multiple past-tense forms appearing in the Corpus of Historical American English[17]. The table shows each lemma, its corresponding regular and irregular forms, the number of times it occurs in the simple past tense in the corpus, and the FIT *P* value for rejecting the neutral null hypothesis. The last three columns show the population size (*N*) and selection coefficient (*s*) of the regular form inferred by maximum likelihood in the two-parameter model (letting *N* and *s* vary), and the inferred population size (*N*) in the one-parameter model in which *s* is set to zero. A positive *s* indicates selection for the regular form (regularization), whereas negative *s* indicates selection against the regular form (irregularization) with strength given by the magnitude of *s*.

Extended Data Table 2 | List of similar-sounding monomorphic verbs for each past-tense conjugation of polymorphic verbs

	Rhyme scheme		Similar-sounding lemmas	
Lemma	Regular	Irregular	Regular	Irregular
light	light/lighted	light/lit	invite, excite, delight, cite	bite
wake	wake/waked	wake/woke	stake, rake, fake, brake, bake, ache	break
sneak	sneak/sneaked	sneak/snuck	wreak, tweak, streak, squeak, shriek, pique, peek, peak, leak, freak, creak	
weave	weave/weaved	weave/wove	retrieve, relieve, receive, perceive, grieve, deceive, conceive, believe, achieve	
dive	dive/dived	dive/dove	thrive, survive, revive, derive, deprive, contrive, arrive	strive, drive
smell	smell/smelled	smell/smelt	yell, swell, shell, repel, rebel, quell, propel, parallel, impel, fell, expel, excel, dispel, compel	
wed	wed/wedded	wed/wed	thread, spearhead, shred, imbed, head, dread, behead	spread, shed
quit	quit/quitted	quit/quit	visit, submit, spirit, solicit, remit, pit, permit, outwit, outfit, forfeit, elicit, deposit, credit, counterfeit, commit, benefit, admit, acquit	split, slit, hit
spell	smell/smelled	smell/smelt	yell, swell, shell, repel, rebel, quell, propel, parallel, impel, fell, expel, excel, dispel, compel	
shine	shine/shined	shine/shone	undermine, underline, twine, sign, resign, refine, redesign, redefine, recline, pine, outline, opine, mine, malign, line, incline, headline, fine, entwine, dine, design, define, decline, consign, confine, combine, assign, align	
tell	smell/smelled	tell/told	yell, swell, shell, repel, rebel, quell, propel, parallel, impel, fell, expel, excel, dispel, compel	sell
leap	leap/leaped	leap/leapt	seep, reap, peep, heap, beep	weep, sweep, sleep, keep, creep
build	build/builded	build/built	gild	rebuild
plead	plead/pleaded	plead/pled	supersede, succeed, stampede, seed, secede, recede, proceed, precede, need, intercede, impede, heed, exceed, concede, cede, bead, accede	mislead, feed, breed, bleed
learn	learn/learned	learn/learnt	yearn, turn, sunburn, spurn, return, overturn, iron, govern, earn, discern, concern, churn, adjourn	
heave	weave/weaved	weave/wove	retrieve, relieve, receive, perceive, grieve, deceive, conceive, believe, achieve	
dwell	smell/smelled	smell/smelt	yell, swell, shell, repel, rebel, quell, propel, parallel, impel, fell, expel, excel, dispel, compel	
wet	wet/wetted	wet/wet	whet, sweat, silhouette, regret, pet, net, jet, fret, abet	upset, set, offset, let, bet, beset
burn	learn/learned	learn/learnt	yearn, turn, sunburn, spurn, return, overturn, iron, govern, earn, discern, concern, churn, adjourn	
kneel	kneel/kneeled	kneel/knelt	wheel, unseal, steel, seal, reveal, repeal, reel, peel, heel, heal, congeal, conceal, appeal	feel, deal
spoil	spoil/spoiled	spoil/spoilt	toil, soil, roil, recoil, oil, foil, coil, broil, boil	
awake	wake/waked	wake/woke	stake, rake, fake, brake, bake, ache	break
know	know/knowed	know/knew	zero, winnow, wallow, veto, tow, torpedo, toe, swallow, stow, sow, snow, slow, show, shadow, sew, row, radio, owe, overshadow, overflow, narrow, mow, hollow, glow, furrow, foreshadow, follow, flow, elbow, echo, crow, bow, borrow, billow, bestow, bellow	overthrow, blow
speed	plead/pleaded	plead/pled	supersede, succeed, stampede, seed, secede, recede, proceed, precede, need, intercede, impede, heed, exceed, concede, cede, bead, accede	mislead, feed, breed, bleed
lay	lay/laid	lay/lay	weigh, waylay, underpay, sway, survey, stray, stay, spray, ricochet, replay, repay, relay, pray, portray, play, pay, parlay, outweigh, obey, flay, display, disobey, dismay, delay, decay, convey, buffet, betray, bay, array, allay	
spill	spill/spilled	spill/spilt	will, thrill, still, stencil, refill, mill, kill, grill, fulfil, fill, drill, distil	
throw	know/knowed	know/knew	zero, winnow, wallow, veto, tow, torpedo, toe, swallow, stow, sow, snow, slow, show, shadow, sew, row, radio, owe, overshadow, overflow, narrow, mow, hollow, glow, furrow, foreshadow, follow, flow, elbow, echo, crow, bow, borrow, billow, bestow, bellow	overthrow, blow
strew	strew/strewed	strew/strew	woo, view, value, undervalue, tattoo, sue, subdue, stew, spew, shoo, screw, review, rescue, renew, pursue, issue, interview, glue, eschew, ensue, discontinue, debut, coo, continue, construe, chew, brew, boo, argue, accrue	
hear	hear/heared	hear/heard	disappear	overhear
knit	knit/knitted	quit/quit	transmit, omit, fit, emit, budget	split, slit, hit
hang	hang/hanged	hang/hung	harangue, bang	
dream	dream/dreamed	dream/dreamt	team, stream, steam, seem, scream, scheme, redeem, gleam, esteem, deem, cream, beam	
catch	catch/catched	catch/caught	snatch, scratch, patch, match, latch, hatch, detach, attach	
lean	lean/leaned	lean/leant	wean, screen, preen, intervene, glean, convene, clean, careen	
draw	draw/drawed	draw/drew	thaw, saw, paw, outlaw, gnaw, claw, awe	withdraw
grow	know/knowed	know/knew	zero, winnow, wallow, veto, tow, torpedo, toe, swallow, stow, sow, snow, slow, show, shadow, sew, row, radio, owe, overshadow, overflow, narrow, mow, hollow, glow, furrow, foreshadow, follow, flow, elbow, echo, crow, bow, borrow, billow, bestow, bellow	overthrow, blow

For each polymorphic verb in this study, its regular and irregular variants each define a present/past rhyme scheme. A monomorphic lemma is considered similar-sounding if both its lemma and its past tense rhyme with a given rhyme scheme. A lemma is included in the table if (1) it has only one simple past-tense form occurring over 50 times in the Corpus of Historical American English, (2) both the past-tense form and the lemma itself are present in the Carnegie Mellon University Pronouncing Dictionary, and (3) the lemma and past-tense forms fit either the regular or irregular rhyme scheme of one of the polymorphic verbs.

Extended Data Table 3 | FIT results for *do*-support

Context	Count	FIT p-value	Inferred N	Inferred s
negative interrogative	606	0.270	504	0.013
affirmative interrogative	4,665	0.178	3,545	0.006
negative declarative	14,227	0.005	2,477	0.014
negative imperative	1,231	0.003	23,150	0.014

We analysed the rise of *do*-support in Early Modern English in four grammatical contexts using instances of potential *do*-support from the Penn Parsed Corpora of Historical English. The table indicates each context, the corresponding number of tokens of potential *do*-support in the corpus, the FIT *P* value for rejecting the neutral null hypothesis, the inferred population size (*N*), and the inferred selection coefficient (*s*) in favour of *do*-support.

natureresearch

Corresponding author(s): Joshua B. Plotkin

□ Initial submission □ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

Sample size was pre-determined by the extent of available data in the digital corpora we analyzed.

2. Data exclusions

Describe any data exclusions.

Exclusions: rare verbs as listed in S1.4, S1.5 and anachronistic documents in S1.6.

3. Replication

Describe whether the experimental findings were reliably reproduced.

There are no experiments in this study.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

There are no experiments in this study.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

There are no experiments in this study.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☒	□	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☒	□	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	□	A statement indicating how many times each experiment was replicated
□	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
□	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
□	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☒	□	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☒	□	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	All software used to analyze the data is freely available, including instructions for reproducibility, at github.com/mnewberry/ldrift

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	n/a

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	n/a

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	n/a
b. Describe the method of cell line authentication used.	n/a
c. Report whether the cell lines were tested for mycoplasma contamination.	n/a
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	n/a

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	n/a

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	n/a