LETTERS

Quantifying the evolutionary dynamics of language

Erez Lieberman[1,2,3]*, Jean-Baptiste Michel[1,4]*, Joe Jackson[1], Tina Tang[1] & Martin A. Nowak[1]

Human language is based on grammatical rules[1–4]. Cultural evolution allows these rules to change over time[5]. Rules compete with each other: as new rules rise to prominence, old ones die away. To quantify the dynamics of language evolution, we studied the regularization of English verbs over the past 1,200 years. Although an elaborate system of productive conjugations existed in English's proto-Germanic ancestor, Modern English uses the dental suffix, '-ed', to signify past tense[6]. Here we describe the emergence of this linguistic rule amidst the evolutionary decay of its exceptions, known to us as irregular verbs. We have generated a data set of verbs whose conjugations have been evolving for more than a millennium, tracking inflectional changes to 177 Old-English irregular verbs. Of these irregular verbs, 145 remained irregular in Middle English and 98 are still irregular today. We study how the rate of regularization depends on the frequency of word usage. The half-life of an irregular verb scales as the square root of its usage frequency: a verb that is 100 times less frequent regularizes 10 times as fast. Our study provides a quantitative analysis of the regularization process by which ancestral forms gradually yield to an emerging linguistic rule.

Natural languages comprise elaborate systems of rules that enable one speaker to communicate with another[7]. These rules serve to simplify the production of language and enable an infinite array of comprehensible formulations[8–10]. However, each rule has exceptions, and even the rules themselves wax and wane over centuries and millennia[11,12]. Verbs that obey standard rules of conjugation in their native language are called regular verbs[13]. In the Modern English language, regular verbs are conjugated into the simple past and past-participle forms by appending the dental suffix '-ed' to the root (for instance, infinitive/simple past/past participle: talk/talked/talked). Irregular verbs obey antiquated rules (sing/sang/sung) or, in some cases, no rule at all (go/went)[14,15]. New verbs entering English universally obey the regular conjugation (google/googled/googled), and many irregular verbs eventually regularize. It is much rarer for regular verbs to become irregular: for every 'sneak' that 'snuck' in[16], there are many more 'flews' that 'flied' out.

Although less than 3% of modern verbs are irregular, the ten most common verbs are all irregular (be, have, do, go, say, can, will, see, take, get). The irregular verbs are heavily biased towards high frequencies of occurrence[17,18]. Linguists have suggested an evolutionary hypothesis underlying the frequency distribution of irregular verbs: uncommon irregular verbs tend to disappear more rapidly because they are less readily learned and more rapidly forgotten[19,20].

To study this phenomenon quantitatively, we studied verb inflection beginning with Old English (the language of *Beowulf*, spoken around AD 800), continuing through Middle English (the language of Chaucer's *Canterbury Tales*, spoken around AD 1200), and ending with Modern English, the language as it is spoken today. The modern '-ed' rule descends from Old English 'weak' conjugation, which applied to three-quarters of all Old English verbs[21]. The exceptions—ancestors of the modern irregular verbs—were mostly members of the so-called 'strong' verbs. There are seven different classes of strong verbs with exemplars among the Modern English irregular verbs, each with distinguishing markers that often include characteristic vowel shifts. Although stable coexistence of multiple rules is one possible outcome of rule dynamics, this is not what occurred in English verb inflection[22]. We therefore define regularity with respect to the modern '-ed' rule, and call all these exceptional forms 'irregular'.

We consulted a large collection of grammar textbooks describing verb inflection in these earlier epochs, and hand-annotated every irregular verb they described (see Supplementary Information). This provided us with a list of irregular verbs from ancestral forms of English. By eliminating verbs that were no longer part of Modern English, we compiled a list of 177 Old English irregular verbs that remain part of the language to this day. Of these 177 Old English irregulars, 145 remained irregular in Middle English and 98 are still irregular in Modern English. Verbs such as 'help', 'grip' and 'laugh', which were once irregular, have become regular with the passing of time.

Next we obtained frequency data for all verbs by using the CELEX corpus, which contains 17.9 million words from a wide variety of textual sources[23]. For each of our 177 verbs, we calculated the frequency of occurrence among all verbs. We subdivided the frequency spectrum into six logarithmically spaced bins from 10^{-6} to 1. Figure 1a shows the number of irregular verbs in each frequency bin. There are only two verbs, 'be' and 'have', in the highest frequency bin, with mean frequency $>10^{-1}$. Both remain irregular to the present day. There are 11 irregular verbs in the second bin, with mean frequency between 10^{-2} and 10^{-1}. These 11 verbs have all remained irregular from Old English to Modern English. In the third bin, with a mean frequency between 10^{-3} and 10^{-2}, we find that 37 irregular verbs of Old English all remained irregular in Middle English, but only 33 of them are irregular in Modern English. Four verbs in this frequency range, 'help', 'reach', 'walk' and 'work', underwent regularization. In the fourth frequency bin, 10^{-4} to 10^{-3}, 65 irregular verbs of Old English have left 57 in Middle and 37 in Modern English. In the fifth frequency bin, 10^{-5} to 10^{-4}, 50 irregulars of Old English have left 29 in Middle and 14 in Modern English. In the sixth frequency bin, 10^{-6} to 10^{-5}, 12 irregulars of Old English decline to 9 in Middle and only 1 in Modern English: 'slink', a verb that aptly describes this quiet process of disappearance (Table 1).

Plotting the number of irregular verbs against their frequency generates a unimodal distribution with a peak between 10^{-4} and 10^{-3}. This unimodal distribution again demonstrates that irregular verbs are not an arbitrary subset of all verbs, because a random subset of verbs (such as all verbs that contain the letter 'm') would follow Zipf's law, a power law with a slope of -0.75 (refs 24,25).

[1]Program for Evolutionary Dynamics, Department of Organismic and Evolutionary Biology, Department of Mathematics, [2]Department of Applied Mathematics, Harvard University, Cambridge, Massachusetts 02138, USA. [3]Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [4]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA.
*These authors contributed equally to this work

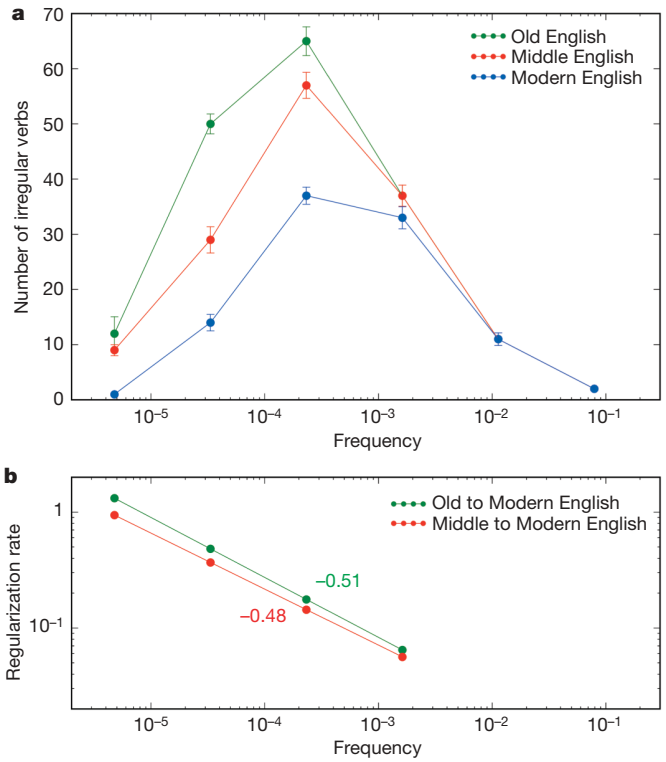

Figure 1 | Irregular verbs regularize at a rate that is inversely proportional to the square root of their usage frequency. **a**, The evolution of 177 verbs from Old English (green) over time, through Middle English (red) and Modern English (blue). The fraction remaining irregular in each bin decreases as the frequency decreases. The frequency shown is that of the modern descendant, and was computed using the CELEX corpus. Error bars indicate standard deviation and were calculated using the bootstrap method. **b**, The regularization rate of irregular verbs as a function of frequency. The relative regularization rates obtained by comparing Old versus Modern English (green) and Middle versus Modern English (red) scale linearly on a log–log plot with a downward slope of nearly one-half. The regularization rate and the half-life scale with the square root of the frequency.

Four of our six frequency bins, those between 10^{-6} and 10^{-2}, allow us to estimate the relative regularization rates of irregular verbs. Calculating the relative regularization rates of verbs of different frequencies is independent of time, which makes the dating of Old and Middle English irrelevant for this calculation. We can plot regularization rate versus frequency and fit a straight line in a log–log plot (Fig. 1b). By comparing Old and Modern English we obtain a slope of about -0.51. Therefore, an irregular verb that is 100 times less frequent is regularized 10 times as fast. In other words, the half-life of irregular verbs is proportional to the square root of their frequency. By comparing Middle and Modern English we find a slope of about -0.48, consistent with the previous result. Both comparisons show that low-frequency irregular verbs are selectively forgotten.

Figure 2a shows the exponential decay of the irregular verbs in the four frequency bins between 10^{-6} and 10^{-2} as a function of time. From these data, which depend on the dating of Old and Middle English, we can estimate actual half-lives of the irregular verbs in different frequency bins. Irregular verbs that occur with a frequency between 10^{-6} and 10^{-5} have a half-life of about 300 years, whereas those with a frequency between 10^{-4} and 10^{-3} have a half-life of 2,000 years. If we fit half-life versus frequency with a straight line in a log–log plot, we obtain a slope of 0.50, which again suggests that the half-life of irregular verbs is proportional to approximately the square root of their frequency (Fig. 2b). It is noteworthy that various methods of fitting the data give the same results.

We cannot directly determine the regularization rate for frequency bins above 10^{-2}, because regularization is so slow that no event was observed in the time span of our data; however, we can extrapolate. For instance, the half-life of verbs with frequencies between 10^{-2} and 10^{-1} should be 14,400 years. For these bins, the population is so small and the half-life so long that we may not see a regularization event in the lifetime of the English language.

To test whether the dynamics within individual competing rules were captured by our global analysis, we studied the decay of individual classes of strong verbs (for example, hit/hit/hit, hurt/hurt/hurt; draw/drew/drawn, grow/grew/grown)[26]. Although our resolution is limited by the small sample size, exponential decay is once again observed, with similar exponents (see Supplementary Fig. 1). Like a Cheshire cat, dying rules vanish one instance at a time, leaving behind a unimodal frown.

Because adequate corpora of Old and Middle English do not exist, we have estimated the frequency of an irregular verb of Old and Middle English by the frequency of the corresponding (regular or irregular) verb of Modern English[27]. A large fraction of verbs would have had to change frequency by several orders of magnitude to

Table 1 | The 177 irregular verbs studied

Frequency	Verbs	Regularization (%)	Half-life (yr)
10^{-1}–1	be, have	0	38,800
10^{-2}–10^{-1}	come, do, find, get, give, go, know, say, see, take, think	0	14,400
10^{-3}–10^{-2}	begin, break, bring, buy, choose, draw, drink, drive, eat, fall, fight, forget, grow, hang, help, hold, leave, let, lie, lose, reach, rise, run, seek, set, shake, sit, sleep, speak, stand, teach, throw, understand, walk, win, work, write	10	5,400
10^{-4}–10^{-3}	arise, bake, bear, beat, bind, bite, blow, bow, burn, burst, carve, chew, climb, cling, creep, dare, dig, drag, flee, float, flow, fly, fold, freeze, grind, leap, lend, lock, melt, reckon, ride, rush, shape, shine, shoot, shrink, sigh, sing, sink, slide, slip, smoke, spin, spring, starve, steal, step, stretch, strike, stroke, suck, swallow, swear, sweep, swim, swing, tear, wake, wash, weave, weep, weigh, wind, yell, yield	43	2,000
10^{-5}–10^{-4}	bark, bellow, bid, blend, braid, brew, cleave, cringe, crow, dive, drip, fare, fret, glide, gnaw, grip, heave, knead, low, milk, mourn, mow, prescribe, redden, reek, row, scrape, seethe, shear, shed, shove, slay, slit, smite, sow, span, spurn, sting, stink, strew, stride, swell, tread, uproot, wade, warp, wax, wield, wring, writhe	72	700
10^{-6}–10^{-5}	bide, chide, delve, flay, hew, rue, shrive, slink, snip, spew, sup, wreak	91	300

177 Old English irregular verbs were compiled for this study. These are arranged according to frequency bin, and in alphabetical order within each bin. Also shown is the percentage of verbs in each bin that have regularized. The half-life is shown in years. Verbs that have regularized are indicated in red. As we move down the list, an increasingly large fraction of the verbs are red; the frequency-dependent regularization of irregular verbs becomes immediately apparent.

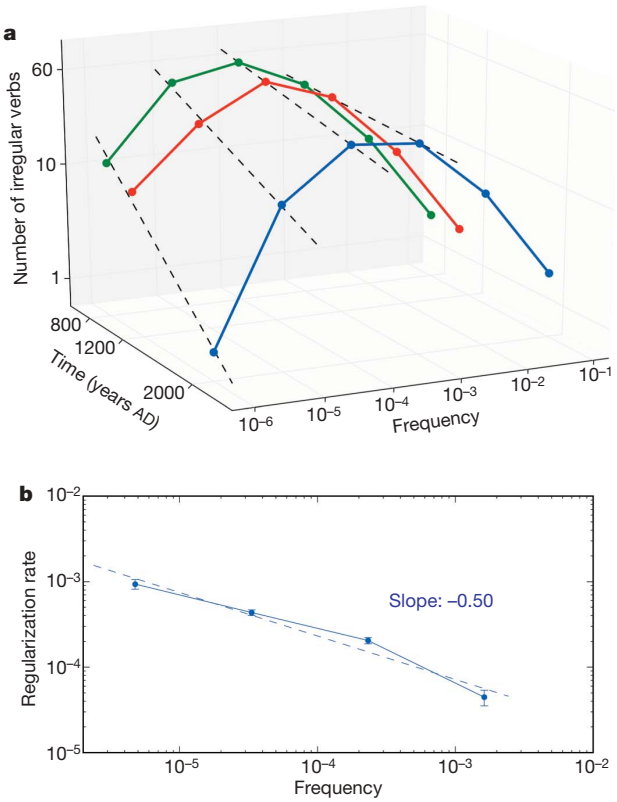

Figure 2 | Irregular verbs decay exponentially over time. **a**, Specifying approximate dates of Old and Middle English allows computation of absolute regularization rates. Regularization rates increase as frequencies decrease, but are otherwise constant over time. **b**, Absolute rates of regularization are shown as a function of frequency. Error bars indicate standard deviation and were calculated using the bootstrap method. The square-root scaling is obtained again.

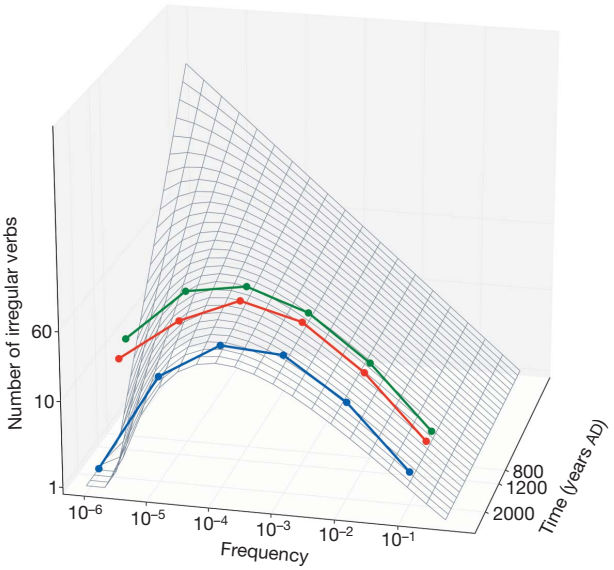

Figure 3 | Extrapolating forward and backward in time using the observation that regularization rate scales as the square root of frequency. The differential system is exactly solvable and the solution fits all three observed distributions. As we move backward in time, the distribution of irregular verbs approaches the zipfian distribution characteristic of random sets of words. The distribution for exceptions to the '-ed' rule became non-random because of frequency-dependent regularization due to selective pressure from the emerging rule.

interfere with the effects observed. To verify that large changes in frequency are rare, we compared frequency data from CELEX with frequencies drawn from the largest available corpus of Middle English texts[28]. Of 50 verbs, only 5 had frequency changes greater than a factor of 10 (Supplementary Fig. 2).

Our analysis covers a vast period, spanning the Norman invasion and the invention of the printing press, but these events did not upset the dynamics of English regularization. Therefore, it is possible to retrospectively trace the evolution of the irregular verbs, moving backwards in time from the observed Modern English distribution and up through Middle and Old English. Going still further back in time allows us to explore the effects of completely undoing the frequency-dependent selective process that the irregular verbs have undergone. Eventually, the shape of the curve changes from unimodal to a power law decline, with slope of nearly −0.75 (Fig. 3). This finding is consistent with the fact that random subsets of verbs (and of all types of words) exhibit such a zipfian distribution. The observed irregular verb distribution is the result of selective pressure on a random collection of ancestral verbs.

We can also make predictions about the future of the past tense. By the time one verb from the set 'begin, break, bring, buy, choose, draw, drink, drive, eat, fall' will regularize, five verbs from the set 'bid, dive, heave, shear, shed, slay, slit, sow, sting, stink' will be regularized. If the current trends continue, only 83 of the 177 verbs studied will be irregular in 2500.

What will be the next irregular verb to regularize? It is likely to be wed/wed/wed. The frequency of 'wed' is only 4.2 uses per million verbs, ranking at the very bottom of the modern irregular verbs. Indeed, it is already being replaced in many contexts by wed/wedded/wedded. Now is your last chance to be a 'newly wed'. The married couples of the future can only hope for 'wedded' bliss.

In previous millennia, many rules vied for control of English language conjugation, and fossils of those rules remain to this day. Yet, from this primordial soup of conjugations, the dental suffix '-ed' emerged triumphant. The competing rules are long dead, and unfamiliar even to well-educated native speakers. These rules disappeared because of the gradual erosion of their instances by a process that we call regularization. But regularity is not the default state of a language—a rule is the tombstone of a thousand exceptions.

METHODS SUMMARY

We searched 11 reference works on Old and Middle English, compiling a list of every irregular verb that we found. We determined whether each verb is still present in Modern English. For all Old English verbs whose descendants remained in the English language, we checked whether they were still irregular using a complete listing of the modern irregular verbs. If they had regularized, we determined when regularization had occurred on the basis of the last time period in which we found a positive annotation listing the verb as irregular. A list of sources used and the entire resulting annotation are provided in the Supplementary Information.

We determined usage frequencies for all the verbs using the CELEX database. We then binned the Old English irregular verbs using a standard logarithmic binning algorithm in Python. We used the resulting binning to determine regularization rates for verbs of differing frequencies. Regularization rates (Fig. 1b) for each bin were computed directly. The fits to exponential decay (Fig. 2) and to the solution of the irregular equation (Fig. 3 and Supplementary Information) were produced using the method of least squares. The Python source code for producing the figures and the table is available at http://www.languagedata.org.

Received 20 March; accepted 27 July 2007.

1. Chomsky, N. *Aspects of the Theory of Syntax* (MIT Press, Cambridge, 1965).
2. Lightfoot, D. *The Development of Language: Acquisition, Change and Evolution* (Blackwell, Oxford, 1999).
3. Clark, R. & Roberts, I. A computational model of language learnability and language change. *Linguist. Inq.* **24,** 299–345 (1993).
4. Abrams, D. & Strogatz, S. Modelling the dynamics of language death. *Nature* **424,** 900 (2003).
5. Nowak, M. A., Komarova, N. L. & Niyogi, P. Computational and evolutionary aspects of language. *Nature* **417,** 611–617 (2002).

6. Hooper, J. in *Current Progress in Historical Linguistics* (ed. Christie, W.) 95–105 (North-Holland, Amsterdam, 1976).
7. Hauser, M. D., Chomsky, N. & Fitch, W. T. The faculty of language: what is it, who has it, and how did it evolve? *Science* **298,** 1569–1579 (2002).
8. Chomsky, N. & Lasnik, H. in *Syntax: An International Handbook of Contemporary Research* (ed. Jacobs, J.) 506–569 (de Gruyte, Berlin, 1993).
9. Dougherty, R. C. *Natural Language Computing* (Lawrence Erlbaum, Hillsdale, 1994).
10. Stabler, E. P. & Keenan, E. L. Structural similarity within and among languages. *Theor. Comput. Sci.* **293,** 345–363 (2003).
11. Niyogi, P. *The Computational Nature of Language Learning and Evolution* (MIT Press, Cambridge, 2006).
12. Labov, W. Transmission and diffusion. *Language* **83,** 344–387 (2007).
13. Pinker, S. *Words and Rules: The Ingredients of Language* (Basic Books, New York, 1999).
14. Kroch, A. Reflexes of grammar in patterns of language change. *Lang. Var. Change* **1,** 199–244 (1989).
15. Kroch, A. in *Papers from the 30th Regional Meeting of the Chicago Linguistics Society: Parasession on Variation and Linguistic Theory* (eds Beals, K. *et al.*) 180–201 (CLS, Chicago, 1994).
16. Pinker, S. The irregular verbs. *Landfall* 83–85 (Autumn issue, 2000).
17. Bybee, J. *Morphology: a Study of Relation Between Meaning and Form* (Benjamins, Amsterdam, 1985).
18. Greenberg, J. in *Current Trends in Linguistics III* (eds Sebeok, T. A. *et al.*) 61–112 (Mouton, The Hague, 1966).
19. Bybee, J. From usage to grammar: the mind's response to repetition. *Language* **82,** 711–733 (2006).
20. Corbett, G., Hippisley, A., Brown, D. & Marriott, P. in *Frequency and the Emergence of Linguistic Structure* (eds Bybee, J. & Hopper, P.) 201–226 (Benjamins, Amsterdam, 2001).
21. Hare, M. & Elman, J. Learning and morphological change. *Cognition* **56,** 61–98 (1995).
22. Marcus, G., Brinkmann, U., Clahsen, H., Wiese, R. & Pinker, S. German inflection: the exception that proves the rule. *Cognit. Psychol.* **29,** 189–256 (1995).
23. Van der Wouden, T. in *Papers from the 3rd International EURALEX Congress* (eds Magay, T. & Zigány, J.) 363–373 (Akadémiai Kiadó, Budapest, 1988).
24. Zipf, G. K. *Human Behavior and the Principle of Least Effort* (Addison-Wesley, Cambridge, 1949).
25. Miller, G. A. Some effects of intermittent silence. *Am. J. Psychol.* **70,** 311–314 (1957).
26. Yang, C. *Knowledge and Learning in Natural Language* (Oxford Univ. Press, New York, 2002).
27. Glushko, M. Towards the quantitative approach to studying evolution of English verb paradigm. *Proc. 19th Scand. Conf. Ling.* **31,** 30–45 (2003).
28. Kroch, A. & Taylor, A. *Penn-Helsinki Parsed Corpus of Middle English* [CD-ROM] 2nd edn (2000) ⟨http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-2/⟩.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the John Templeton Foundation and by a grant from the NSF-NIH joint programme in mathematical biology. The Program for Evolutionary Dynamics is sponsored by J. Epstein. E.L. was supported by the National Defense Science and Engineering Graduate Fellowship and the National Science Foundation Graduate Fellowship. We thank S. Pinker, J. Rau, D. Donoghue and A. Presser for discussions, and J. Saragosti for help with visualization.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to E.L. (erez@erez.com).