The structure and evolution of language
Principles of Complex Systems
CSYS/MATH 300, Fall, 2011

Prof. Peter Dodds

Department of Mathematics & Statistics | Center for Complex Systems | Vermont Advanced Computing Center | University of Vermont

Licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License.
Outline

Irregular verbs

Word lifespans

Meanings

References
Irregular verbs

Cleaning up English:

“Quantifying the evolutionary dynamics of language” [1]

▶ Exploration of how verbs with irregular conjugation gradually become regular over time.
▶ Comparison of verb behavior in Old, Middle, and Modern English.
Irregular verbs

- Universal tendency towards regular conjugation
- Rare verbs tend to be regular in the first place
Rates are relative.

The more common a verb is, the more resilient it is to change.
Rates are relative.

The more common a verb is, the more resilient it is to change.
Four of our six frequency bins, those between 10^26 and 10^22, allow us to estimate the relative regularization rates of verbs of different frequency bins. Irregular verbs are selectively regularized by a factor of 10 to 100 times faster than regular verbs. Thus, for very low-frequency verbs, the half-life of the regular verbs is proportional to approximately the square root of their frequency (Fig. 2b). It is noteworthy that various individual classes of strong verbs (for example, hit/hit/hit, hurt/hurt/hurt) have had to change frequency by several orders of magnitude to achieve regularity.

We cannot directly determine the regularization rate for frequency bins above 10^2. By comparing Middle and Modern English we find a slope of about 0.50, which again suggests that the half-life of irregular verbs is proportional to approximately the square root of their frequency (Fig. 2b). The regularization relative regularization rates obtained by comparing Old versus Modern English indicate standard deviation and were calculated using the bootstrap method. The fraction remaining irregular in each bin from Old English (green) over time, through Middle English (red) and to the square root of their usage frequency. a Calculating the relative regularization rates of verbs of different frequency, we can estimate actual half-lives of the irregular verbs in Old to Modern English. The evolution of 177 verbs is regularized 10 times as fast. In other words, the half-life of the English language.

### Table 1 | The 177 irregular verbs studied

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Verbs</th>
<th>Regularization (%)</th>
<th>Half-life (yr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10^{-1}–1</td>
<td>be, have</td>
<td>0</td>
<td>38,800</td>
</tr>
<tr>
<td>10^{-2}–10^{-1}</td>
<td>come, do, find, get, give, go, know, say, see, take, think</td>
<td>0</td>
<td>14,400</td>
</tr>
<tr>
<td>10^{-3}–10^{-2}</td>
<td>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</td>
<td>10</td>
<td>5,400</td>
</tr>
<tr>
<td>10^{-4}–10^{-3}</td>
<td>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</td>
<td>43</td>
<td>2,000</td>
</tr>
<tr>
<td>10^{-5}–10^{-4}</td>
<td>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, silt, smite, sow, span, spurn, sting, stink, strewn, stride, swell, tread, uproot, wade, warp, wax, wield, wring, writhe</td>
<td>72</td>
<td>700</td>
</tr>
<tr>
<td>10^{-6}–10^{-5}</td>
<td>bide, chide, delve, flay, hew, rue, shrive, slink, snap, spew, sup, weak</td>
<td>91</td>
<td>300</td>
</tr>
</tbody>
</table>

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.

- **Red** = regularized
- **Estimates of half-life for regularization.**
Irregular verbs

- ‘Wed’ is next to go.
- -ed is the winning rule...

During Middle English, an equation was derived to predict how verbs will be regularized...
Irregular verbs

Regularization rate $\propto$ word frequency$^{-1/2}$

Half life $\propto$ word frequency$^{1/2}$
Irregular verbs

- Projecting back in time...

The evolution of the irregular verbs, moving up through Middle and Old English. Going still further back in time allows us to explore the effects of completely undoing the frequency-dependent regularization. Therefore, it is possible to retroactively trace the evolution of the irregular verbs, moving forward through Modern English. What will be the next irregular verb to regularize? It is likely to be one of the verbs that are currently in the process of regularization. By analyzing the frequency of each verb over the past millennium, we can estimate the time it will take for each verb to regularize. 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 will become more and more regular, as the selective pressure on irregular verbs decreases. The human touch is required to produce the figures and the table is available at http://www.languagedata.org.
Word meanings

Preliminary findings on word frequency and number of meanings

▶ Corpus: 10,000 most frequent words from Project Gutenberg
▶ # meanings for each word estimated using dictionary.com
▶ Friends: perl, regular expressions, wget.
Word meanings

A. Word frequency versus rank, slope $\alpha \sim -1.2$ corresponds to a frequency distribution with $\gamma \sim 1.8$.

B. Relationship between average number of meanings and average frequency (bins are by rank, with each circle representing 500 words). Slope of 1/3 lower than Zipf’s 1/2 \[^4\].
Word meanings

Meaning number as a function of word rank.

The three exponents combine within error:

\[ 1.2 \times \frac{1}{3} = 0.4 \approx 0.45. \]
Scaling collapse for meaning number distribution
Each curve corresponds to approximately 500 words group according to rank (1–500, 501–1000, ...).
With normalization

\[ P(n_m) = f^{-1/3} G \left( f^{-1/3} n_m \right). \]
Word meanings

Further work:

- Check these scalings again
- Explore alternate data sources
- Think about why meaning number might scale with frequency.
- May be an information theoretic story.
- If we add context, we may be able to use a modified version of Simon’s approach \[^3\]
- The city story here would be that there may be many cities and towns with the same name (e.g., Springfield) with an uneven distribution in populations.
Word meanings

Further work:

- Check these scalings again
- Explore alternate data sources
  - Think about why meaning number might scale with frequency.
  - May be an information theoretic story.
  - If we add context, we may be able to use a modified version of Simon’s approach[^3]
  - The city story here would be that there may be many cities and towns with the same name (e.g., Springfield) with an uneven distribution in populations.

[^3]: Referenced work or study.
Word meanings

Further work:

- Check these scalings again
- Explore alternate data sources
- Think about why meaning number might scale with frequency.
- May be an information theoretic story.
- If we add context, we may be able to use a modified version of Simon’s approach [3]
- The city story here would be that there may be many cities and towns with the same name (e.g., Springfield) with an uneven distribution in populations.
Further work:

▶ Check these scalings again
▶ Explore alternate data sources
▶ Think about why meaning number might scale with frequency.
▶ May be an information theoretic story.
▶ If we add context, we may be able to use a modified version of Simon’s approach.\(^3\)
▶ The city story here would be that there may be many cities and towns with the same name (e.g., Springfield) with an uneven distribution in populations.
Word meanings

Further work:

▶ Check these scalings again
▶ Explore alternate data sources
▶ Think about why meaning number might scale with frequency.
▶ May be an information theoretic story.
▶ If we add context, we may be able to use a modified version of Simon’s approach [3]
▶ The city story here would be that there may be many cities and towns with the same name (e.g., Springfield) with an uneven distribution in populations.
Word meanings

Further work:

- Check these scalings again
- Explore alternate data sources
- Think about why meaning number might scale with frequency.
- May be an information theoretic story.
- If we add context, we may be able to use a modified version of Simon’s approach \[3\]
- The city story here would be that there may be many cities and towns with the same name (e.g., Springfield) with an uneven distribution in populations.
References

Quantifying the evolutionary dynamics of language. 

Quantitative analysis of culture using millions of digitized books. 

On a class of skew distribution functions. 