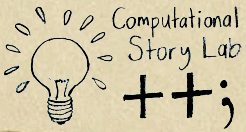
The structure and evolution of language

Principles of Complex Systems | @pocsvox
CSYS/MATH 300, Fall, 2013 | #FallPoCS2013

Prof. Peter Dodds | @peterdodds

Dept. of Mathematics & Statistics | Vermont Complex Systems Center
Vermont Advanced Computing Core | University of Vermont



These slides are brought to you by:

The UNIVERSITY of VERMONT

Outline

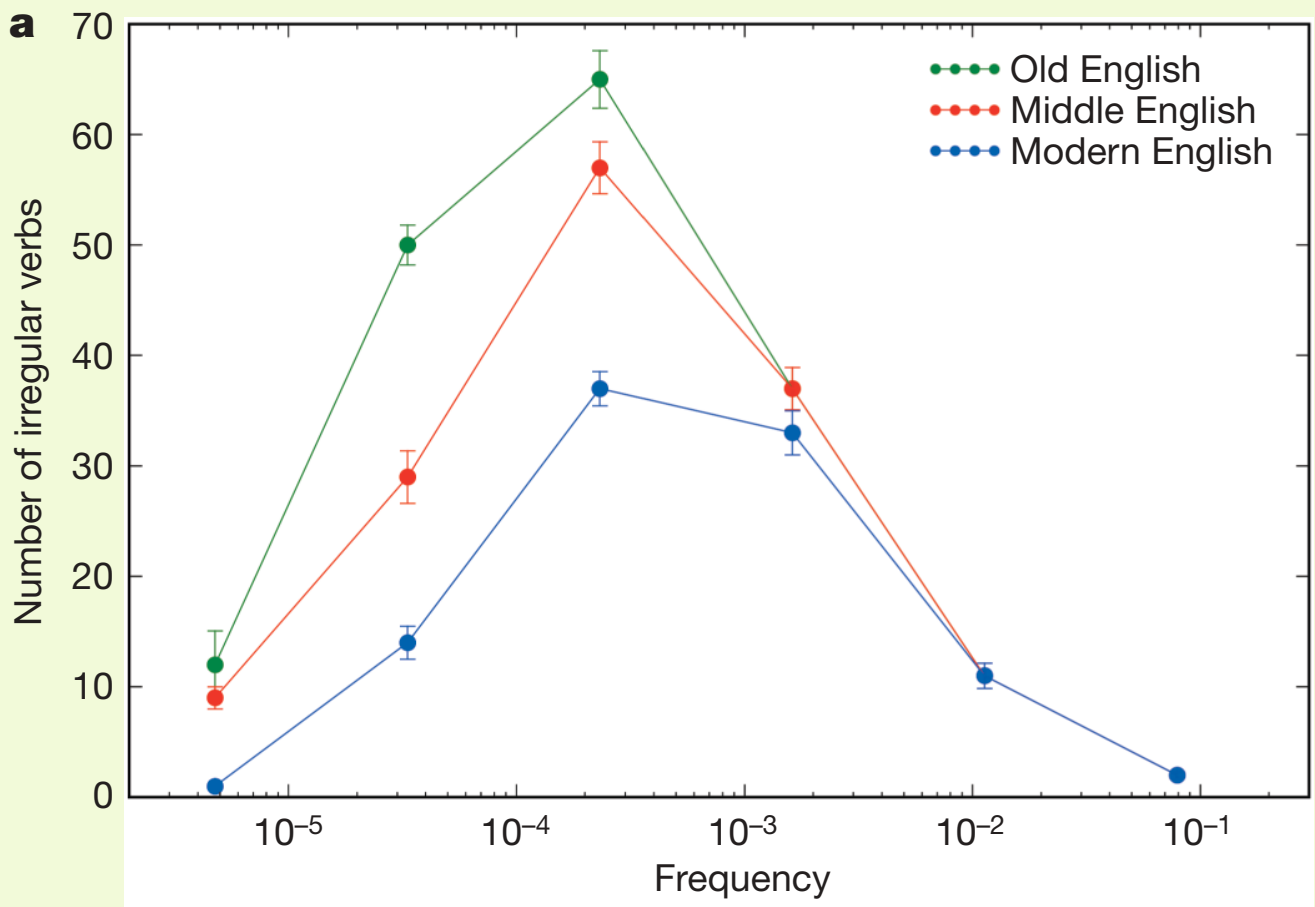
Irregular verbs

Cleaning up English:

"Quantifying the evolutionary dynamics of language" [1]
Lieberman et al., Nature, Vol 449, 713-716, 2007.

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

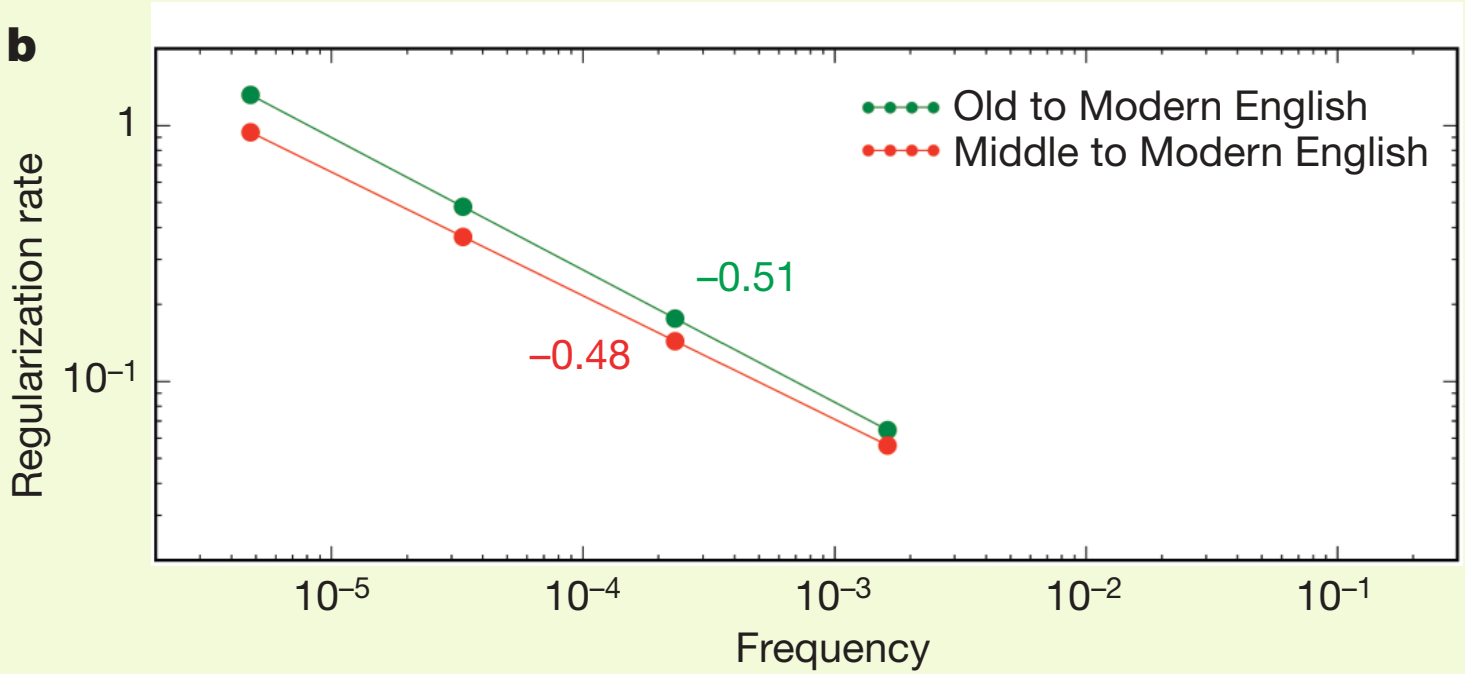
Irregular verbs



- Universal tendency towards regular conjugation
- Rare verbs tend to be regular in the first place

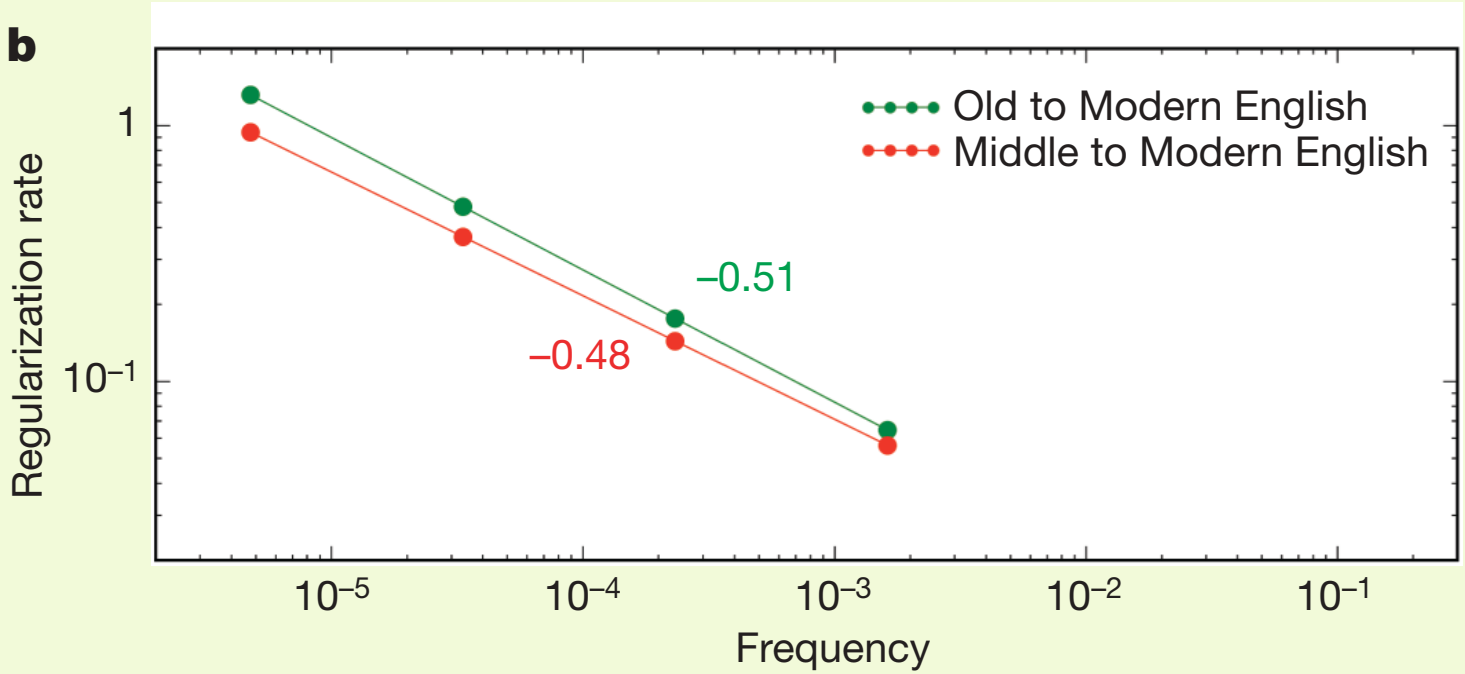
The University of Vermont

Irregular verbs



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

Irregular verbs



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

The UNIVERSITY of VERMONT

Irregular verbs

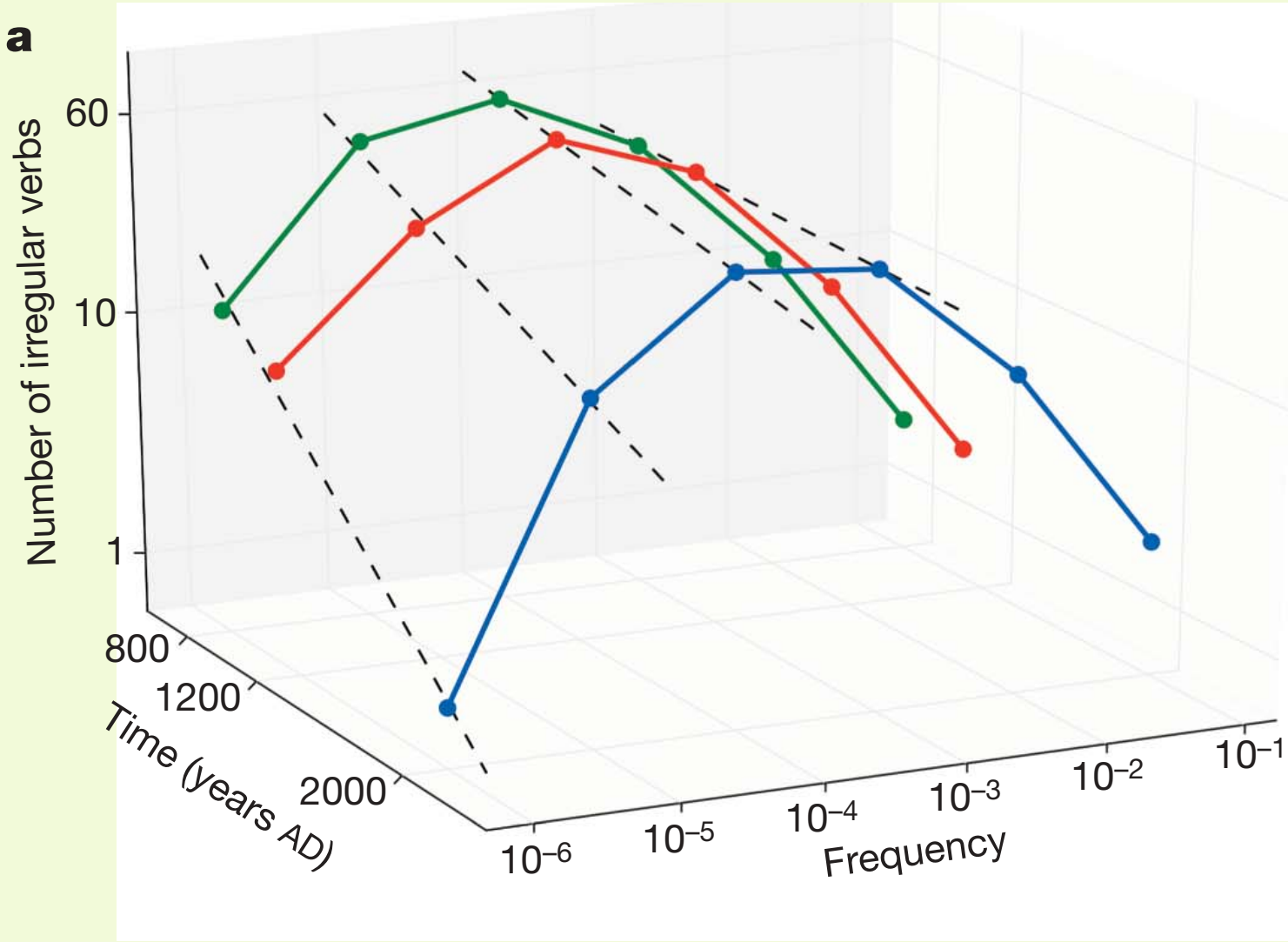
Table 1 | The 177 irregular verbs studied

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.

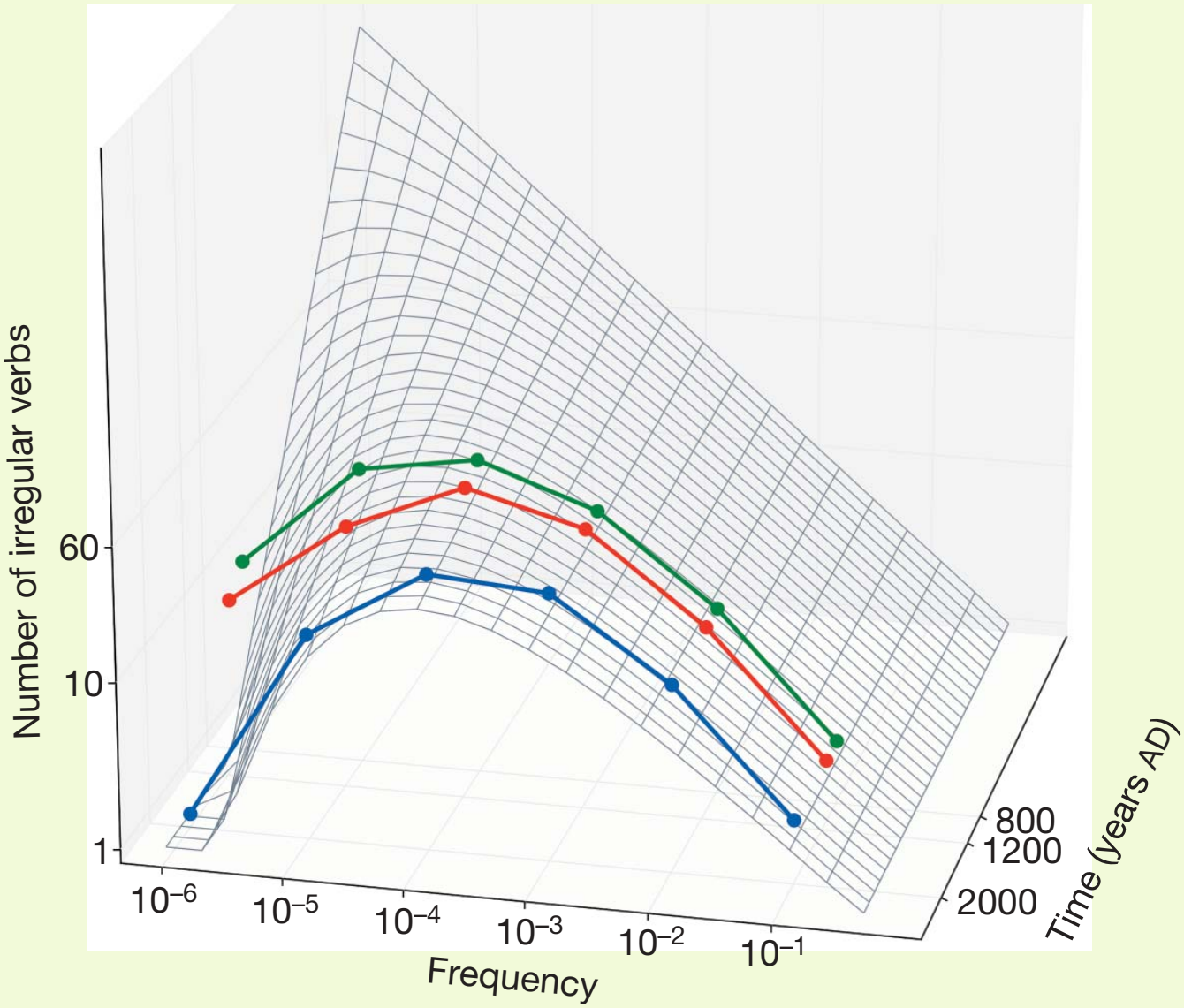
- Red = regularized
- Estimates of half-life for regularization ($\propto f^{1/2}$)

Irregular verbs



- 'Wed' is next to go.
- -ed is the winning rule...

Irregular verbs



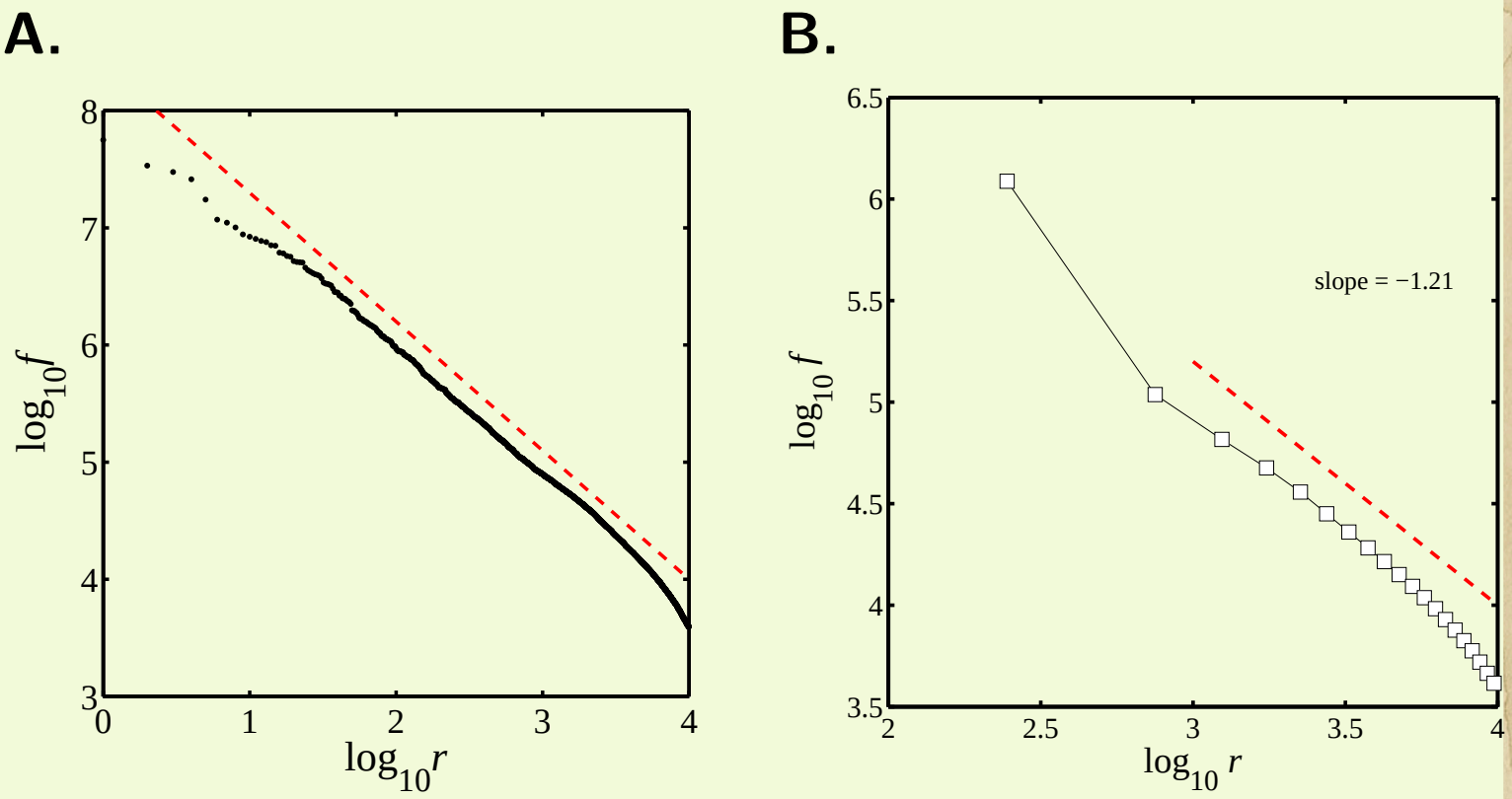
- Projecting back in time to proto-Zipf story of many tools

Word meanings

Preliminary findings on word frequency and number of meanings

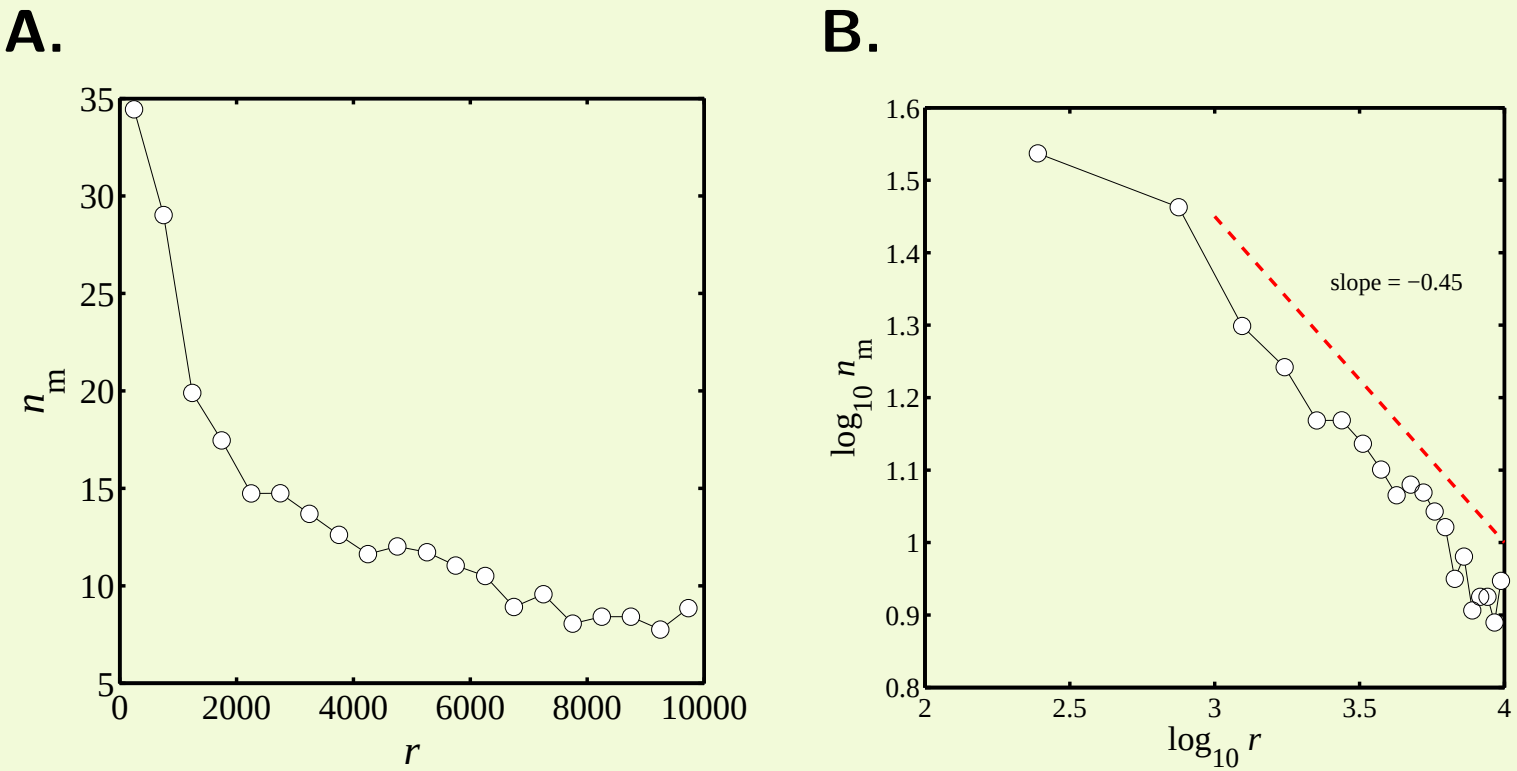
- Corpus: 10,000 most frequent words from Project Gutenberg
- # meanings for each word estimated using dictionary.com
- Friends: perl, regular expressions, wget.

Word meanings



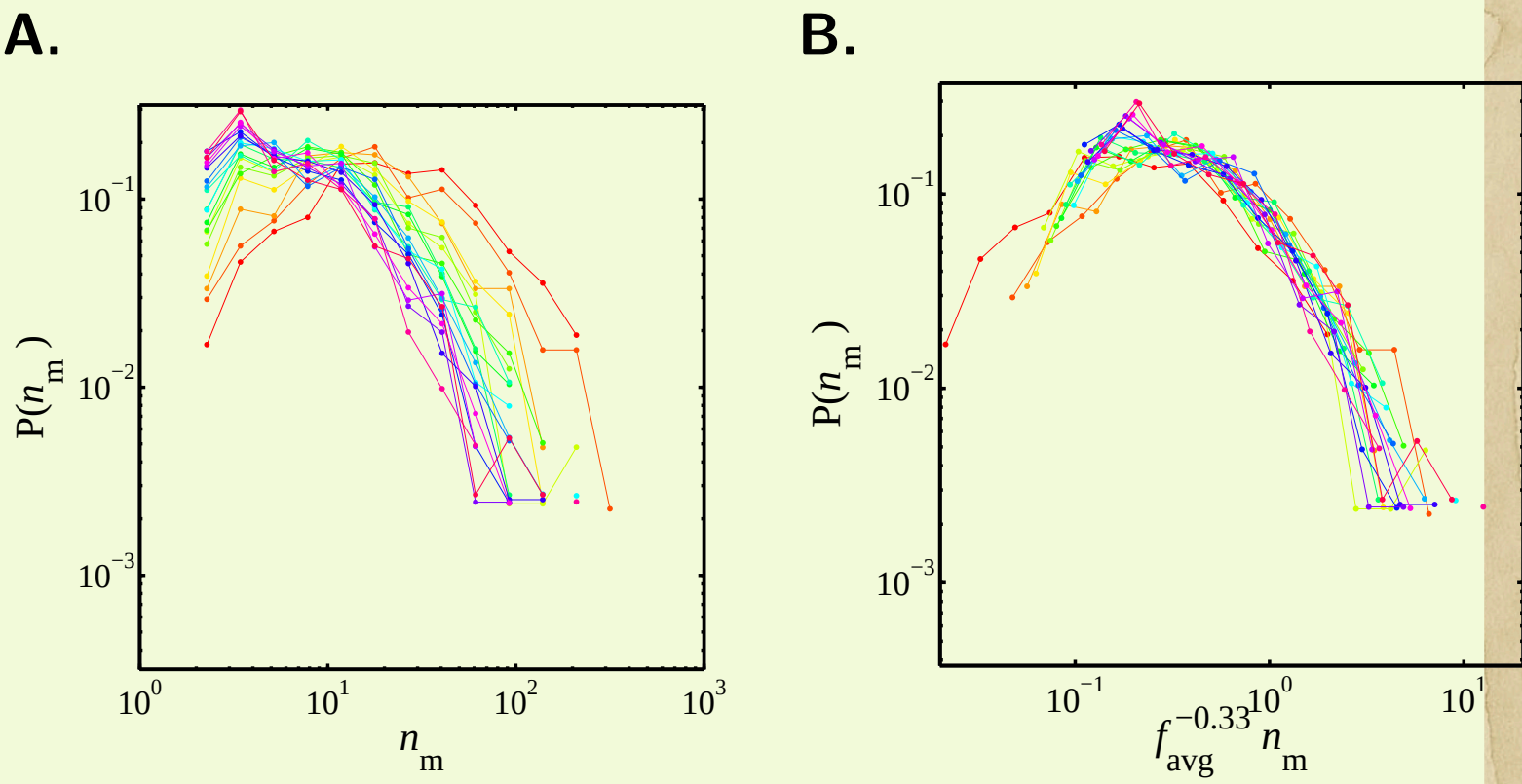
A. Word frequency versus rank, slope $\alpha \sim -1.2$ corresponds to to a frequency distribution with $\gamma \sim 1.8$.
B. Relationship between average number of meanings and average frequency (bins are by rank, with each circle representing 500 words). Slope of 1/3 lower than Zipf's 1/2 [4].

Word meanings



- Meaning number as a function of word rank.
- The three exponents combine within error: $1.2 \times 1/3 = 0.4 \simeq 0.45$.

Word meanings



- 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.

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.

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 I

[1] E. Lieberman, J.-B. Michel, J. Jackson, T. Tang, and M. A. Nowak.
Quantifying the evolutionary dynamics of language.
Nature, 449:713–716, 2007. pdf

[2] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, T. G. B. Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. A. Lieberman.
Quantitative analysis of culture using millions of digitized books.
Science Magazine, 2010. pdf

[3] H. A. Simon.
On a class of skew distribution functions.
Biometrika, 42:425–440, 1955. pdf

The UNIVERSITY of VERMONT

References II

[4] G. K. Zipf.
Human Behaviour and the Principle of Least-Effort.
Addison-Wesley, Cambridge, MA, 1949.