Quantifying Language Evolution with Lexical Turbulence

Last updated: 2022/08/29, 00:04:32 EDT

Principles of Complex Systems, Vols. 1, 2, & 3D
CSYS/MATH 300, 303, & 394, 2022–2023 | @pocsvox

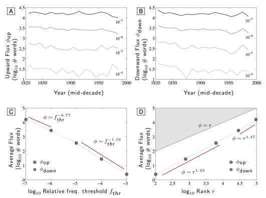
Prof. Peter Sheridan Dodds | @peterdodds

Computational Story Lab | Vermont Complex Systems Center
Santa Fe Institute | University of Vermont

Outline

References

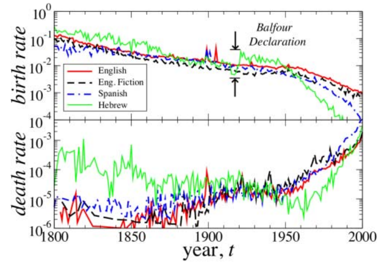
"Is language evolution grinding to a halt? The scaling of lexical turbulence in English fiction suggests it is not"
Pechenick, Danforth, and Dodds.
Journal of Computational Science, **21**, 24–37, 2017. [1]



- Upshot: Not dead yet.

A bit of a worry—language is slowing down:

"Statistical laws governing fluctuations in word use from word birth to word death"
Petersen et al.,
Scientific Reports, **2**, 313, 2012. [2]



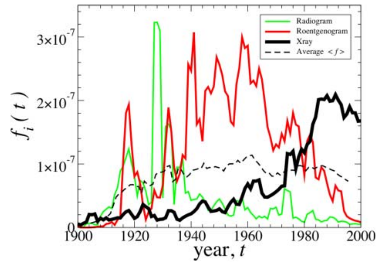


Figure 1 | Word extinction. The English word "Roentgenogram" derives from the Nobel prize winning scientist and discoverer of the X-ray, Wilhelm Röntgen (1845–1923). The prevalence of this word was quickly challenged by two main competitors, "X-ray" (recorded as "Xray" in the database) and "Radiogram." The arithmetic mean frequency of these three time series is relatively constant over the 80-year period 1920–2000, $\langle f \rangle \approx 10^{-7}$, illustrating the limited linguistic "market share" that can be achieved by any competitor. We conjecture that the main reason "Xray" has a higher frequency is due to the "fitness gain" from its efficient short word length and also due to the fact that English has become the base language for scientific publication.

Petersen *et al.* define the birth year and death year of an individual word as the first and last year, respectively, that the given word's relative frequency $f_{w;y}$ is found to be equal to or greater than a cutoff frequency $f^{\text{cut}}_{w;y_1,y_2}$ equal to one twentieth its median relative frequency $f^{\text{med}}_{w;y_1,y_2}$:

$$f_{w;y} \geq f^{\text{cut}}_{w;y_1,y_2} = 0.05 f^{\text{med}}_{w;y_1,y_2}.$$

- y_1 and y_2 = the first and last year of the overall time period.
- Excluded: words appearing in only one year (this turns out to be a problem) and words appearing for the first time before $y_1 = 1700$.
- Rates of word birth and death found by normalizing the numbers of word births and deaths by the total number of unique words in a given year.

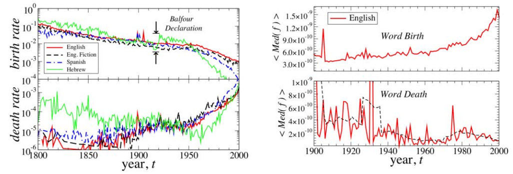


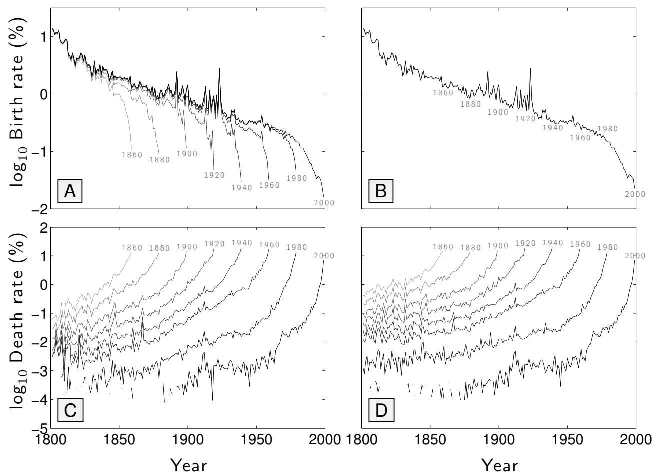
Figure 2 | Dramatic shift in the birth rate and death rate of words. The word birth rate $\gamma_b(t)$ and the word death rate $\gamma_d(t)$ show marked underlying changes in word use competition which affects the entry rate and the sustainability of existing words. The modern print era shows a marked increase in the death rate of words which likely correspond to low fitness, misspelled and (technologically) outdated words. A simultaneous decrease in the birth rate of new words is consistent with the decreasing marginal need for new words indicated by the sub-linear allometric scaling between vocabulary size and total corpus size (Heaps' law)[24]. Interestingly, we quantitatively observe the impact of the Balfour Declaration in 1917, the circumstances surrounding which effectively rejuvenated Hebrew as a national language, resulting in a 5-fold increase in the birth rate of words in the Hebrew corpus.

Figure 3 | Survival of the fittest in the entry process of words. Trends in the relative uses of words that either were born or died in a given year show that the entry-exit forces largely depend on the relative use of the word. For the English corpus, we calculate the average of the median lifetime relative use, $\langle \mathrm{Med}(f_i) \rangle$, for all words born in year t (top panel) and for all words that died in year t (bottom panel), which shows a 5-year moving average (dashed black line). There is a dramatic increase in the relative use ("utility") of newborn words over the last 20–30 years, likely corresponding to new technical terms, which are necessary for the communication of core modern technology and ideas. Conversely, with higher editorial standards and the recent use of word processors which include spelling standardization technology, the words that are dying are those words with low relative use. We confirm by visual inspection that the lists of dying words contain mostly misspelled and nonsensical words.

- Petersen et al. present a range of other interesting observations—all worth looking at [2]
- Our focus will be on life and death of words.

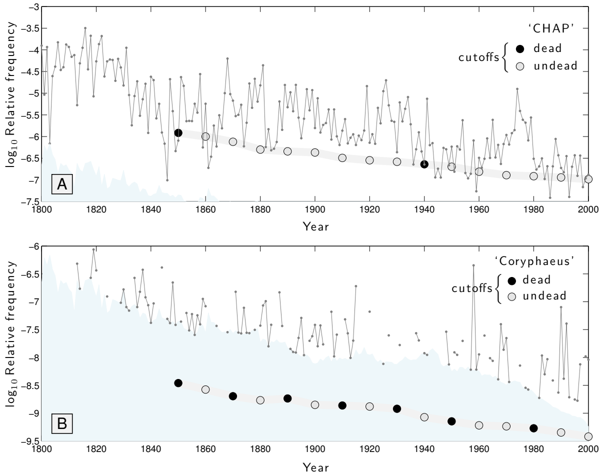
For following pages:

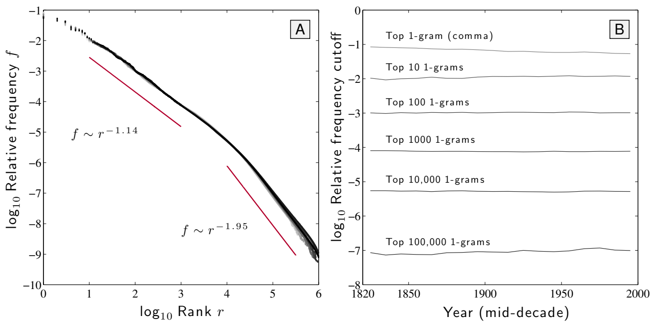
- **A** and **C**: Birth and death rates for 1-grams for the 2012 version of English Fiction determined using the method of Petersen *et al.* [2].
- Curves correspond to different end-of-history boundaries with history running from y_1=1800 to y_2=1860 to 2000 in 20 year increments.
- Birth rates show clear departures from an overall form as each end of history year is approached.
- Including words that appear in only one year in a time range eliminates these discrepancies (plot **B**).
- Death rates however are strongly affected by the choice of when history ends and this cannot be remedied by modifying the rule for 1-gram death.
- As the end of history moves forward in time, words that seemed dead are no longer dead for a number of reasons.
- **B** and **D**: Birth and death rates as per plots **A** and **C** in all respects except now including words that appear only once in a time range—i.e., have a non-zero relative frequency in only one year.
- Birth rates are now well determined retrospectively from any vantage point of history and an exponential decay appears



Why?

- Following: Two examples of how a 1-gram may be variously labeled dead or alive depending on the end of history using the criterion in [2].
- **A.** The word 'CHAP' declines in relative frequency over time, from a high of $10^{-3.5}$ to as low as $10^{-7.5}$.
- Using a twentieth of the median frequency of a 1-gram as a threshold for birth and death, we see 'CHAP' appears to have "run down the curtain" in 1850 but then re-emerged as alive for 8 subsequent decadal end points.
- 'CHAP' once again succumbs in 1940 only to stagger on through 2000.
- This dead-undead cycling can be seen for many words and leads us to exploring how words pass above and drop below fixed relative frequency thresholds.
- In both plots, the blue region marks the lowest possible relative frequency for each year achieved when a 1-gram has a count of 1. **B.**
- The word 'Coryphaeus' is a much less frequent word than 'CHAP', and its time series contains a substantial number of zeroes and ones (resting on the top of the blue region).
- The criterion in [2] leads to a flipping back and forth between





Lexical turbulence:

Zipf's law has two scaling regimes: [3]

$$f \sim \begin{cases} r^{-\alpha} \text{ for } r \ll r_b, \\ r^{-\alpha'} \text{ for } r \gg r_b, \end{cases}$$

When comparing two texts, define Lexical turbulence as flux of words across a frequency threshold:

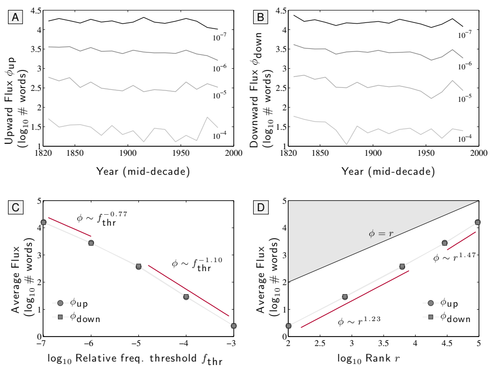
$$\phi \sim \begin{cases} f_{\text{thr}}^{-\mu} \text{ for } f_{\text{thr}} \ll f_b, \\ f_{\text{thr}}^{-\mu'} \text{ for } f_{\text{thr}} \gg f_b, \end{cases}$$

Estimates: $\mu \simeq 0.77$ and $\mu' \simeq 1.10$, and f_b is the scaling break point.

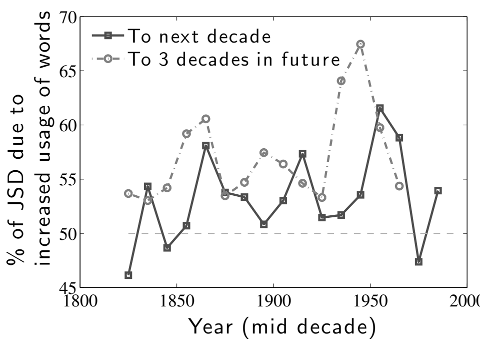
$$\phi \sim \begin{cases} r^{\nu} = r^{\alpha\mu'} \text{ for } r \ll r_b, \\ r^{\nu'} = r^{\alpha'\mu} \text{ for } r \gg r_b. \end{cases}$$

Estimates: Lower and upper exponents $\nu \simeq 1.23$ and $\nu' \simeq 1.47$.

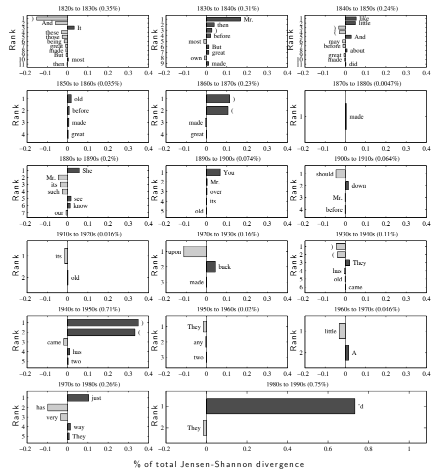
Exponents match up:

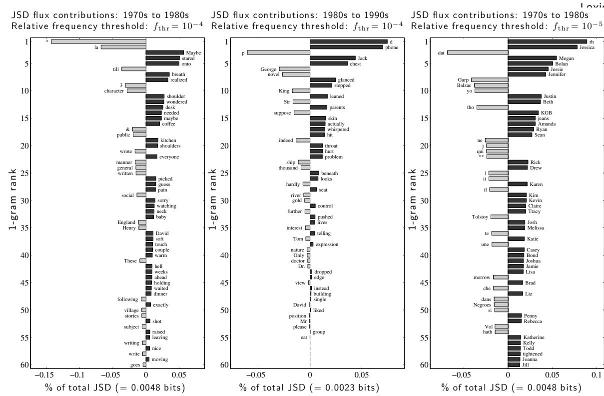


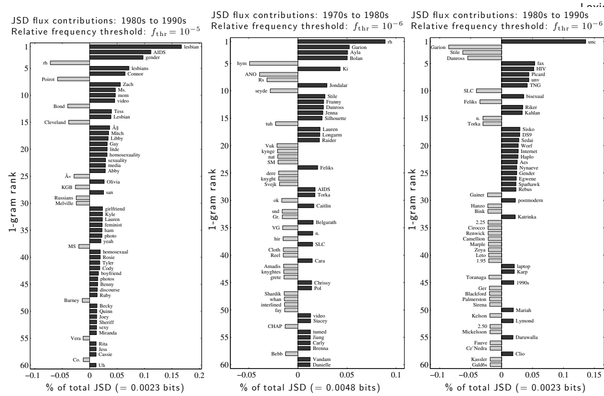
Inter-decade JSD comparisons:

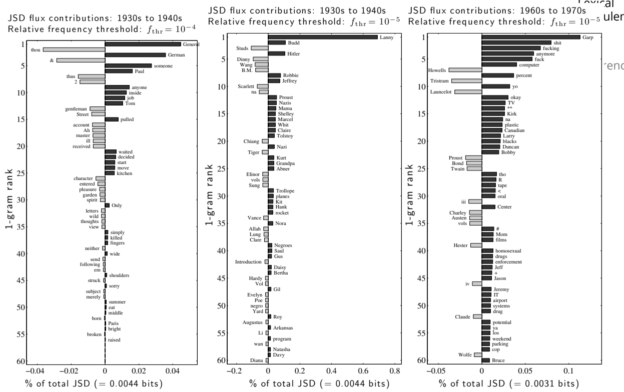


Note: Have moved beyond JSD to rank-turbulence divergence and probability-turbulence divergence.









References I

[1] E. A. Pechenick, C. M. Danforth, and P. S. Dodds.
Is language evolution grinding to a halt? The scaling of lexical turbulence in English fiction suggests it is not.
Journal of Computational Science, 21:24–37, 2017. pdf

[2] A. M. Petersen, J. Tenenbaum, S. Havlin, and H. E. Stanley.
Statistical laws governing fluctuations in word use from word birth to word death.
Scientific Reports, 2:313, 2012. pdf

References II

[3] J. R. Williams, J. P. Bagrow, C. M. Danforth, and P. S. Dodds.
Text mixing shapes the anatomy of rank-frequency distributions.
Physical Review E, 91:052811, 2015. pdf