

# How Google Books misrepresents socio-cultural-linguistic evolution

Last updated: 2024/12/16, 14:01:23 EST

Principles of Complex Systems, Vols. 1, 2, & 3D  
CSYS/MATH 6701, 6713, & a pretend number, 2024–2025

Prof. Peter Sheridan Dodds

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



Licensed under the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/)

These slides are brought to you by:

The PoCVerse  
Corporal Concerns  
2 of 33

Google Books

When Corpora Go Wrong

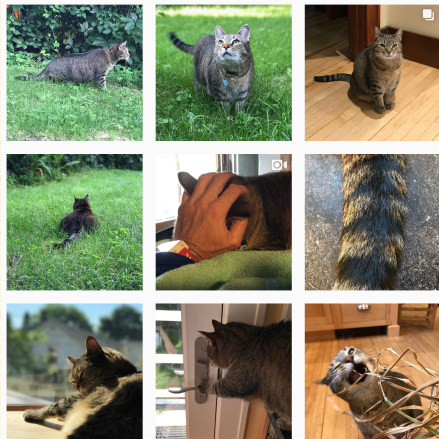
References



Sealie & Lambie  
Productions



These slides are also brought to you by:

Special Guest Executive Producer



 On Instagram at [pratchett\\_the\\_cat](https://www.instagram.com/pratchett_the_cat) 

The PoCverse  
Corporal Concerns  
3 of 33

Google Books

When Corpora Go Wrong

References



# Outline

The PoCverse  
Corporal Concerns  
4 of 33

Google Books

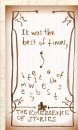
When Corpora Go Wrong

References

Google Books

When Corpora Go Wrong

References

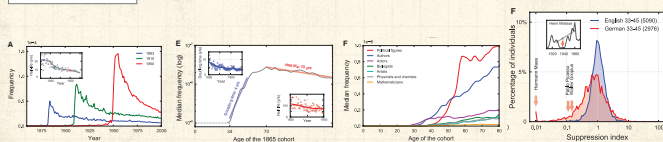




“Quantitative analysis of culture using millions of digitized books” ↗

Michel et al.,

Science Magazine, 331, 176–182, 2011. [1]



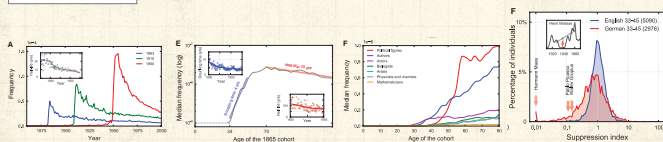
<http://www.culturomics.org/> ↗ and Google Books ngram viewer ↗





“Quantitative analysis of culture using millions of digitized books”

Michel et al.,  
Science Magazine, **331**, 176–182, 2011. [1]



<http://www.culturomics.org/> and Google Books ngram viewer

## Barney Rubble:



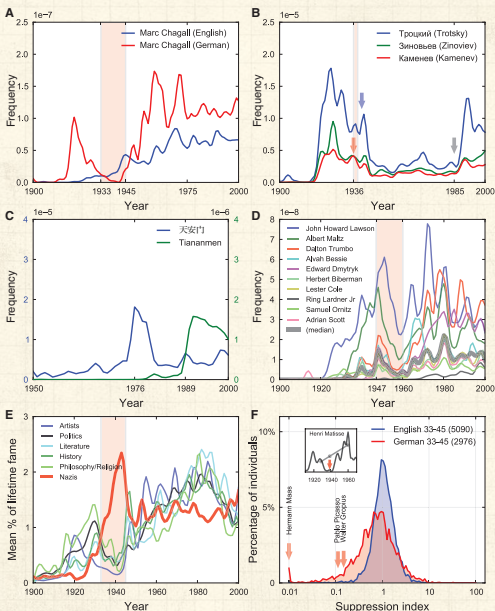
“Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution”

Pechenick, Danforth, and Dodds,  
PLoS ONE, **10**, e0137041, 2015. [2]

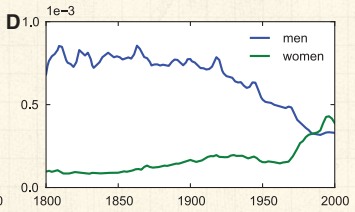
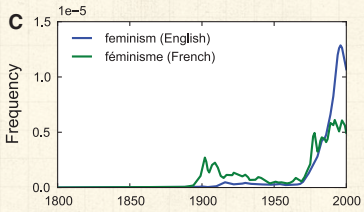
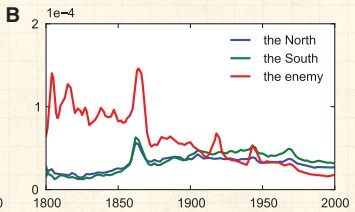
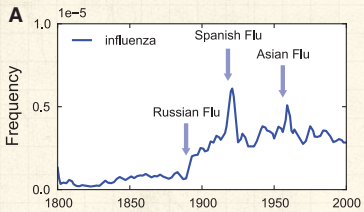



# Censorship (okayish)

**Fig. 4.** Culturomics can be used to detect censorship. **(A)** Usage frequency of "Marc Chagall" in German (red) as compared to English (blue). **(B)** Suppression of Leon Trotsky (blue), Grigory Zinoviev (green), and Lev Kamenev (red) in Russian texts, with noteworthy events indicated: Trotsky's assassination (blue arrow), Zinoviev and Kamenev executed (red arrow), the Great Purge (red highlight), and perestroika (gray arrow). **(C)** The 1976 and 1989 Tiananmen Square incidents both led to elevated discussion in English texts (scale shown on the right). Response to the 1989 incident is largely absent in Chinese texts (blue, scale shown on the left), suggesting government censorship. **(D)** While the Hollywood Ten were blacklisted (red highlight) from U.S. movie studios, their fame declined (median: thick gray line). None of them were credited in a film until 1960's (aptly named *Exodus*). **(E)** Artists and writers in various disciplines were suppressed by the Nazi regime (red highlight). In contrast, the Nazis themselves (thick red line) exhibited a strong fame peak during the war years. **(F)** Distribution of suppression indices for both English (blue) and German (red) for the period from 1933–1945. Three victims of Nazi suppression are highlighted at left (red arrows). Inset: Calculation of the suppression index for "Henri Matisse".



# Danger Will Robinson

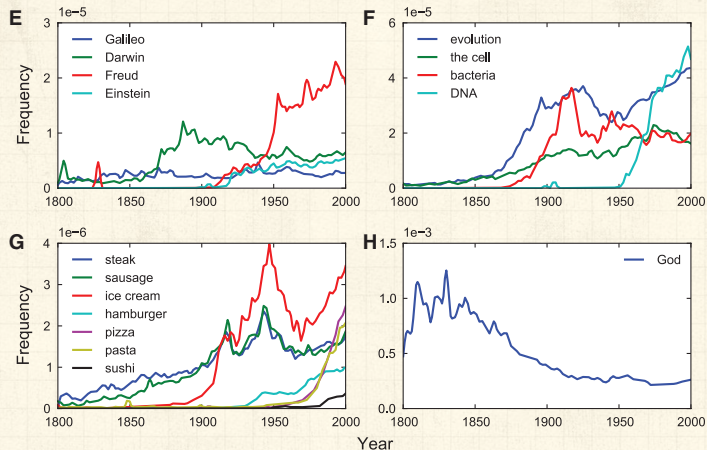


(Search for “cherrypicking” )





# Seriously, Danger Will Robinson



**Fig. 5.** Culturomics provides quantitative evidence for scholars in many fields. (A) Historical epidemiology: “influenza” is shown in blue; the Russian, Spanish, and Asian flu epidemics are highlighted. (B) History of the Civil War. (C) Comparative history. (D) Gender studies. (E and F) History of science. (G) Historical gastronomy. (H) History of religion: “God”.



# Outline

The PoCverse  
Corporal Concerns  
9 of 33

Google Books

When Corpora Go Wrong

References

## Google Books When Corpora Go Wrong

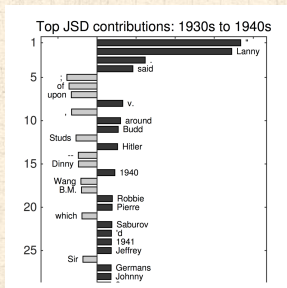
References





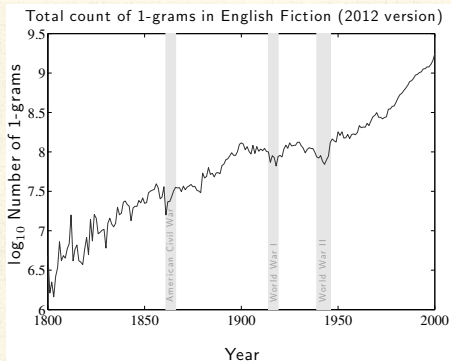
“Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution” ↗

Pechenick, Danforth, and Dodds,  
PLoS ONE, **10**, e0137041, 2015. [2]





## Volume of “words”—exponential growth

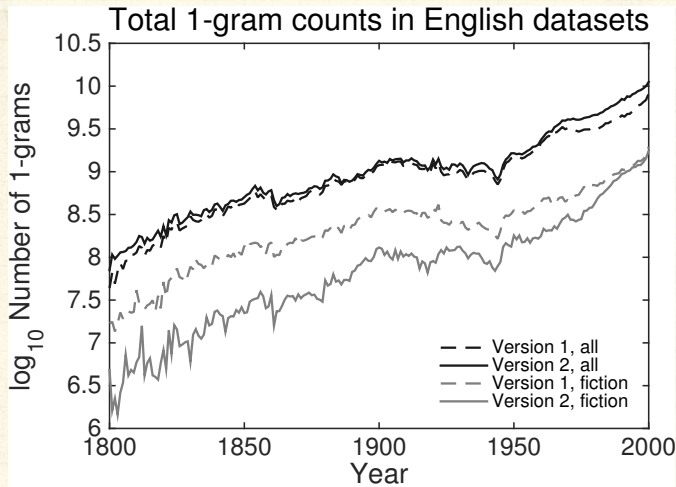


Two data sets: Version 1 (2009, around 4% of all books published) and Version 2 (2012)




Initial version: Around 4% of all published books.

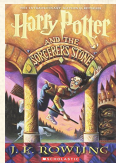






# Trouble at Mill, 1/2:



Every book gets one vote:


 Equally important:



“Harry Potter and the Sorcerer’s Stone”    
by J. K. Rowling (1998). <sup>[3]</sup>



“Microwave Cooking for One”    
by Marie Smith (1999). <sup>[4]</sup>

 New editions, revisions, reprintings give very modest bump.



# Trouble at Mill, 2/2:

## Lord of the Rings is fading away:



Search for Frodo, Gandalf in English Fiction, 2012.



English Fiction = fiction + literary criticism.











## Kullback-Leibler divergence:

Given two distributions  $P$  and  $Q$  over  $N$  categories (e.g., 1-grams):

$$D_{KL}(P||Q) = \sum_{i=1}^N p_i \log_2 \frac{p_i}{q_i},$$




-  Average number of extra bits required to encode a system with true distribution  $P$  under the belief that the true distribution is  $Q$ .
-  Not symmetric.



## Kullback-Leibler divergence:

Given two distributions  $P$  and  $Q$  over  $N$  categories (e.g., 1-grams):

$$D_{KL}(P||Q) = \sum_{i=1}^N p_i \log_2 \frac{p_i}{q_i},$$

-  Average number of extra bits required to encode a system with true distribution  $P$  under the belief that the true distribution is  $Q$ .
-  Not symmetric.
-  Can go kablooey—happens if any  $q_i = 0$ .



## Jensen-Shannon divergence:


$$D_{JS}(P||Q) = \frac{1}{2} (D_{KL}(P||M) + D_{KL}(Q||M)),$$

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

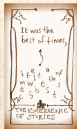


## Jensen-Shannon divergence:

$$D_{JS}(P||Q) = \frac{1}{2} (D_{KL}(P||M) + D_{KL}(Q||M)),$$



  $M = \frac{1}{2}(P + Q)$  is the mixed distribution of  $P$  and  $Q$ .

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



## Jensen-Shannon divergence:

$$D_{JS}(P||Q) = \frac{1}{2} (D_{KL}(P||M) + D_{KL}(Q||M)),$$


-   $M = \frac{1}{2}(P + Q)$  is the mixed distribution of  $P$  and  $Q$ .
-  Symmetric, finite, square root is a metric.


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




## Jensen-Shannon divergence:

$$D_{\text{JS}}(P||Q) = \frac{1}{2} (D_{\text{KL}}(P||M) + D_{\text{KL}}(Q||M)),$$

  $M = \frac{1}{2}(P + Q)$  is the mixed distribution of  $P$  and  $Q$ .

 Symmetric, finite, square root is a metric.

 Rewrite:

$$D_{\text{JS}}(P||Q) = H(M) - \frac{1}{2} (H(P) + H(Q))$$


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







## Jensen-Shannon divergence:


$$D_{JS}(P||Q) = \frac{1}{2} (D_{KL}(P||M) + D_{KL}(Q||M)),$$

  $M = \frac{1}{2}(P + Q)$  is the mixed distribution of  $P$  and  $Q$ .

 Symmetric, finite, square root is a metric.

 Rewrite:

$$D_{JS}(P||Q) = H(M) - \frac{1}{2} (H(P) + H(Q))$$

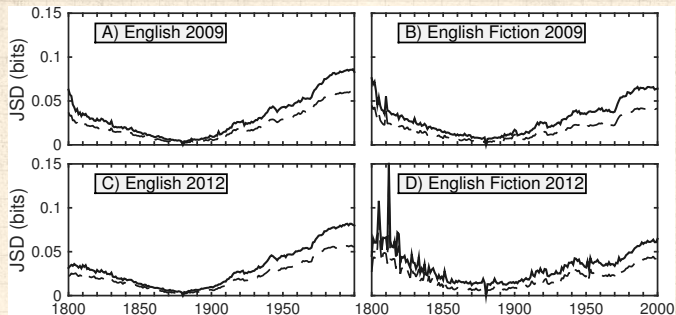
 Use per word contribution to the JSD to make shifts:

$$D_{JS,i}(P||Q) = -m_i \log_2 m_i + \frac{1}{2} (p_i \log_2 p_i + q_i \log_2 q_i)$$

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



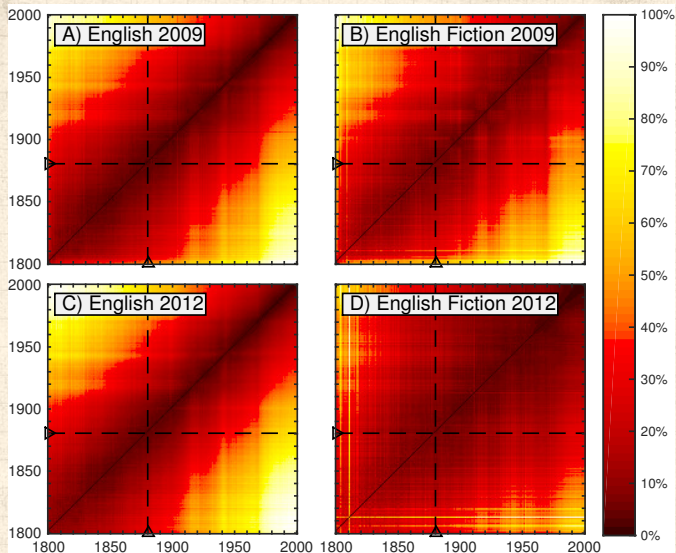
# JSD between 1880 and 1800–2000:



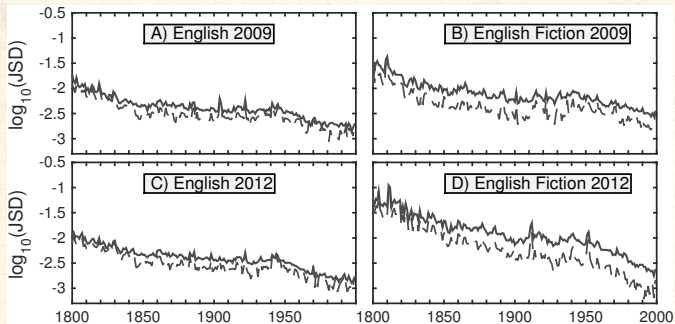
Contributions are counted for all words appearing above a  $10^{-5}$  threshold in a given year; for the dashed curves, the threshold is  $10^{-4}$ .



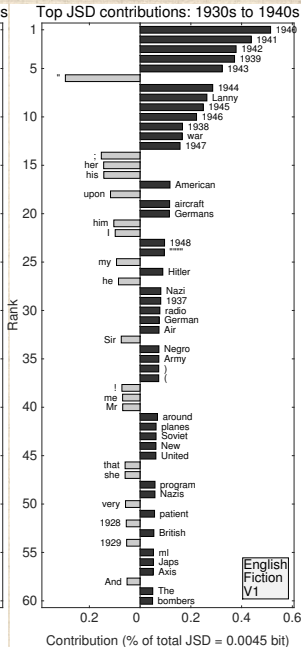
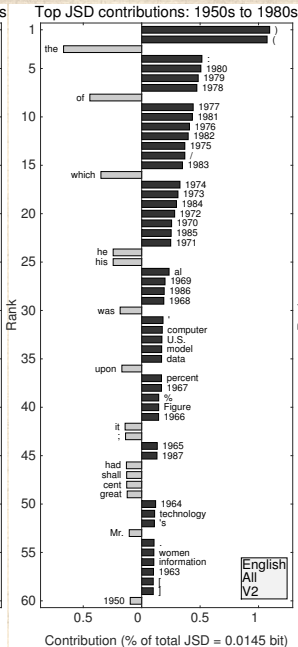
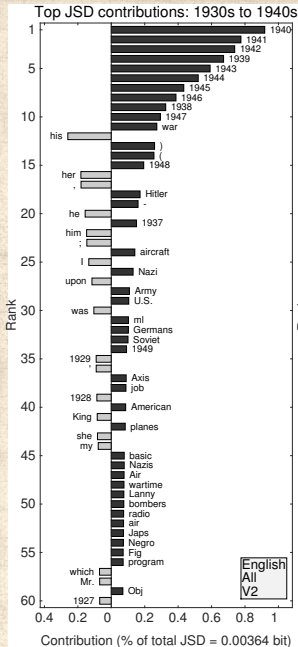
# JSD between years:



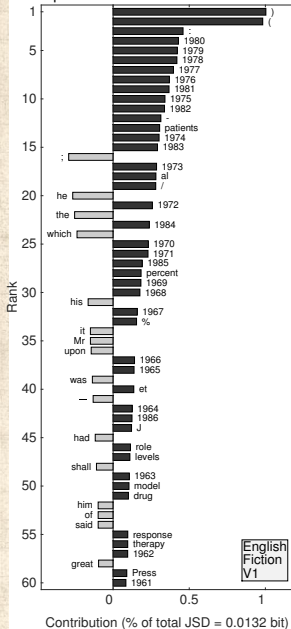
## JSD between consecutive years:



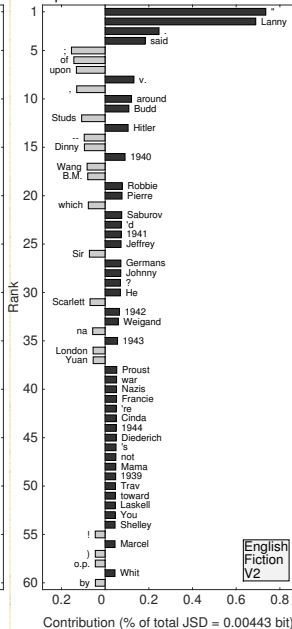
Consecutive year (between each year and the following year) base-10 logarithms of JSD, corresponding to off-diagonals. For the solid curves, contributions are counted for all words appearing above a  $10^{-5}$  threshold in a given year; for the dashed curves, the threshold is  $10^{-4}$ . Divergences between consecutive years typically decline through the mid-19th century, remain relatively steady until the mid-20th century, then continue to decline gradually over time.



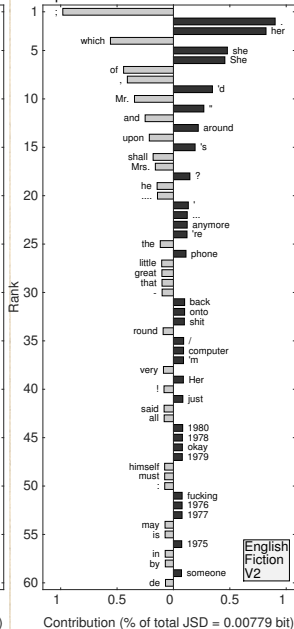
Top JSD contributions: 1950s to 1980s



Top JSD contributions: 1930s to 1940s



Top JSD contributions: 1950s to 1980s



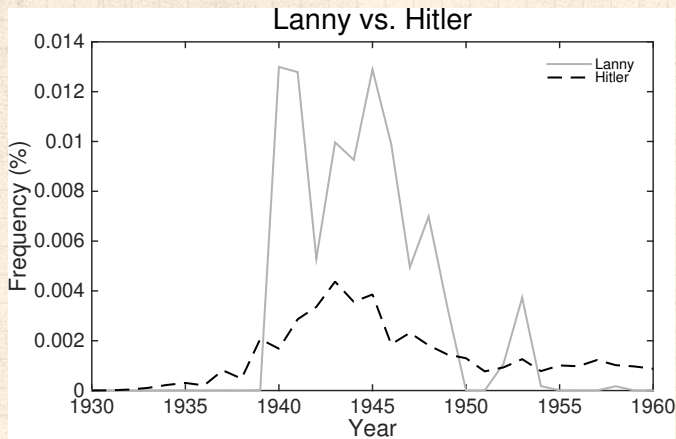
# Lanny Budd, Upton Sinclair's forgotten hero

The PoCVerse  
Corporal Concerns  
24 of 33

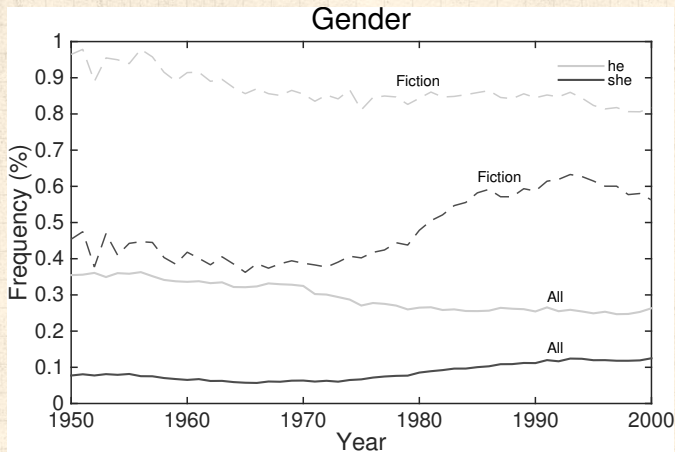
Google Books

When Corpora Go Wrong

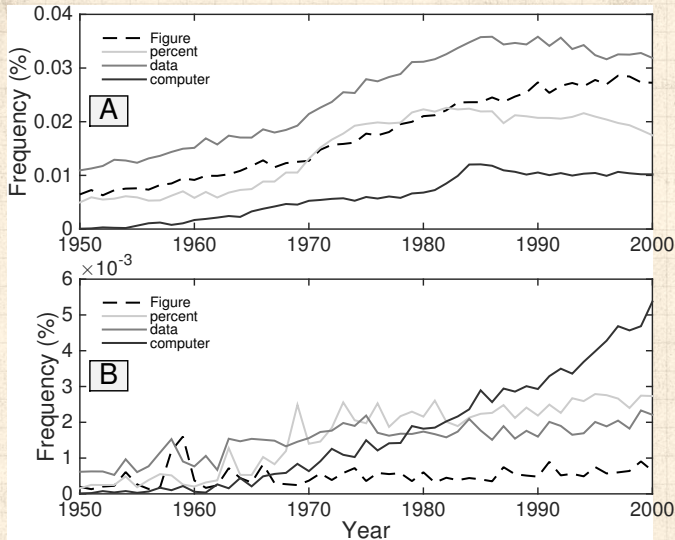
References



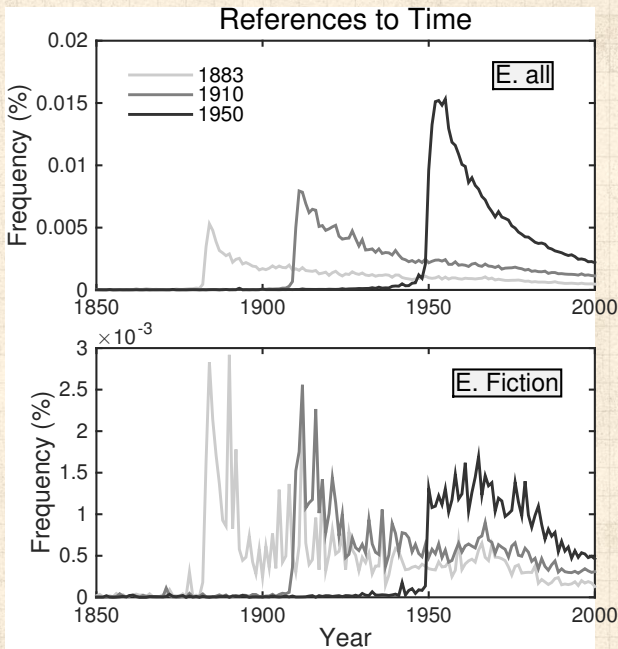
# Representative of a more general shift:







# Science drives the memory story:



# “God is dying” — Google Books



A deeper look reveals that the decline in sacred speech is not a recent trend, though we are only now becoming fully aware of it. By searching the Google Ngram corpus — a collection of millions of books, newspapers, webpages and speeches published between 1500 and 2008 — we can now determine the frequency of word usage over the centuries. This data shows that most religious and spiritual words have been declining in the English-speaking world since the early 20th century.

One might expect a meaty theological term like “salvation” to fade, but basic moral and religious words are also falling out of use. A study in [The Journal of Positive Psychology](#) analyzed 50 terms associated with moral virtue. Language about the virtues Christians call the fruit of the spirit — words like “love,” “patience,” “gentleness” and “faithfulness” — has become much rarer. Humility words, like “modesty,” fell by 52 percent. Compassion words, like “kindness,” dropped by 56 percent. Gratitude words, like “thankfulness,” declined by 49 percent.



[nytimes.com/2018/10/13/opinion/sunday/talk-god-spirituality-christian.html](https://www.nytimes.com/2018/10/13/opinion/sunday/talk-god-spirituality-christian.html)


[theweek.com/articles/791795/death-sacred-speech](https://www.theweek.com/articles/791795/death-sacred-speech) (2018-09-10)

The book to sell: [Learning to Speak God from Scratch: Why Sacred Words Are Vanishing—and How We Can Revive Them](#)

“God feels fine!” —Also Google Books 

Language Log goodness:


 [Lexico-cultural decay?](#) 

<http://languagelog ldc.upenn.edu/nll/?p=40222> 

Mark Liberman

Architecture would appear to be failing with relative decreases in: stairway, foundation, roof, eaves, arch, cornice.


 [“More on trends in the Google ngrams corpus”](#) 

<http://languagelog ldc.upenn.edu/nll/?p=40349> 

Mark Liberman, again

“God talk” words have all been going up after 2000.

We fight the good fight with a (towering) Twitter thread, an essential tool of science:

<https://twitter.com/compstorylab/status/1052708929795497990> 


## Criticism [\[ edit \]](#)

The data set has been criticized for its reliance upon inaccurate **OCR**, an overabundance of scientific literature, and for including large numbers of incorrectly dated and categorized texts.<sup>[12][13]</sup> Because of these errors, and because it is uncontrolled for bias<sup>[14]</sup> (such as the increasing amount of scientific literature, which causes other terms to appear to decline in popularity), it is risky to use this corpus to study language or test theories.<sup>[15]</sup> Since the data set does not include **metadata**, it may not reflect general linguistic or cultural change<sup>[16]</sup> and can only hint at such an effect.

Another issue is that the corpus is in effect a library, containing one of each book. A single, prolific author is thereby able to noticeably insert new phrases into the Google Books lexicon, whether the author is widely read or not.<sup>[14]</sup>

## OCR issues [\[ edit \]](#)

Optical character recognition, or OCR, is not always reliable, and some characters may not be scanned correctly. In particular, systemic errors like the confusion of "s" and "f" in pre-19th century texts (due to the use of the **long\_s** which was similar in appearance to "f") can cause systemic bias. Although Google Ngram Viewer claims that the results are reliable from 1800 onwards, poor OCR and insufficient data mean that frequencies given for languages such as Chinese may only be accurate from 1970 onward, with earlier parts of the corpus showing no results at all for common terms, and data for some years containing more than 50% noise.<sup>[17][18]</sup>

 Ref. 14 = Pechenick *et al.* <sup>[2]</sup>



# Shell of the nut:

The PoCverse  
Corporal Concerns  
31 of 33

Google Books

When Corpora Go Wrong

References



First issue: Google Books has the appearance of cultural popularity.




# Shell of the nut:


The PoCverse  
Corporal Concerns  
31 of 33

Google Books

When Corpora Go Wrong

References

 First issue: Google Books has the appearance of cultural popularity.

 But it's really a representation of a quasi-lexicon.



# Shell of the nut:

- First issue: Google Books has the appearance of cultural popularity.
- But it's really a representation of a quasi-lexicon.
- Depopularizing: Each book appears once (in principle).





# Shell of the nut:

- First issue: Google Books has the appearance of cultural popularity.
- But it's really a representation of a quasi-lexicon.
- Depopularizing: Each book appears once (in principle).
- But natural unevenness of Zipf distribution for words gives veneer of popularity.



# Shell of the nut:

- First issue: Google Books has the appearance of cultural popularity.
- But it's really a representation of a quasi-lexicon.
- Depopularizing: Each book appears once (in principle).
- But natural unevenness of Zipf distribution for words gives veneer of popularity.
- Second issue: Inclusion of massive amounts of scientific literature makes a mess.



# Shell of the nut:

- First issue: Google Books has the appearance of cultural popularity.
- But it's really a representation of a quasi-lexicon.
- Depopularizing: Each book appears once (in principle).
- But natural unevenness of Zipf distribution for words gives veneer of popularity.
- Second issue: Inclusion of massive amounts of scientific literature makes a mess.
- Upshot: Google Books needs a lot more metadata.



# References I

- [1] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books 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](#), 331:176–182, 2011. pdf ↗

- [2] E. A. Pechenick, C. M. Danforth, and P. S. Dodds.  
Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution.

[PLoS ONE](#), 10:e0137041, 2015. pdf ↗

- [3] J. K. Rowling.  
[Harry Potter and the Sorcerer's Stone](#).  
Scholastic Press, New York, 1998.



# References II

The PoCverse  
Corporal Concerns  
33 of 33

Google Books

When Corpora Go Wrong

References

- [4] M. Smith.  
Microwave Cooking for One.  
Pelican Publishing, 1999.

