How Google Books misrepresents socio-cultural-linguistic evolution

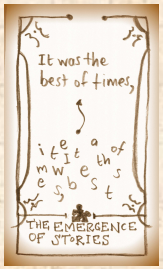
Last updated: 2019/03/13, 21:34:59

Complex Networks | @networksvox
CSYS/MATH 303, Spring, 2019

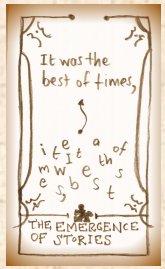
Prof. Peter Dodds | @peterdodds

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



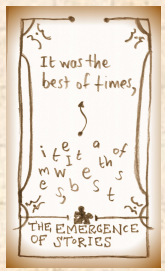
These slides are brought to you by:



These slides are also brought to you by:

Special Guest Executive Producer

On Instagram at pratchett_the_cat

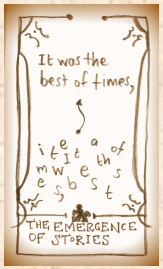


Outline

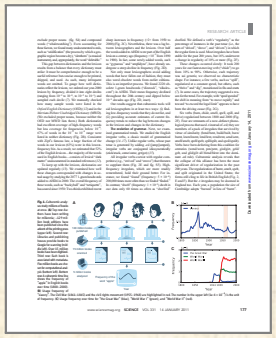
Google Books

When Corpora Go Wrong

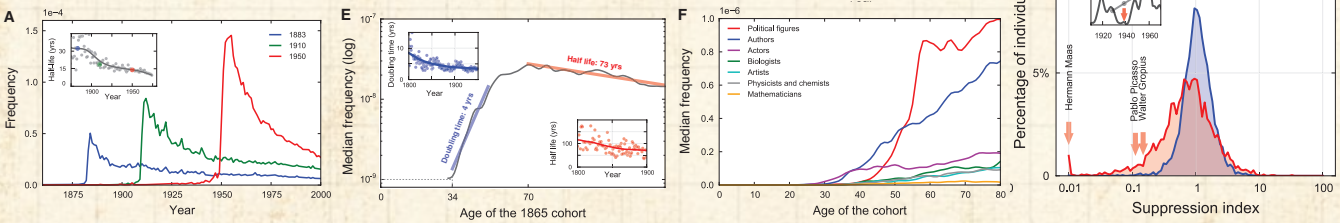
References



Culturomics:



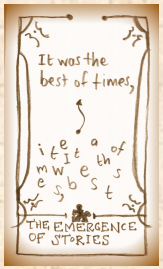
"Quantitative analysis of culture using millions of digitized books"
Michel et al.,
Science Magazine, **331**, 176–182, 2011. [1]



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)

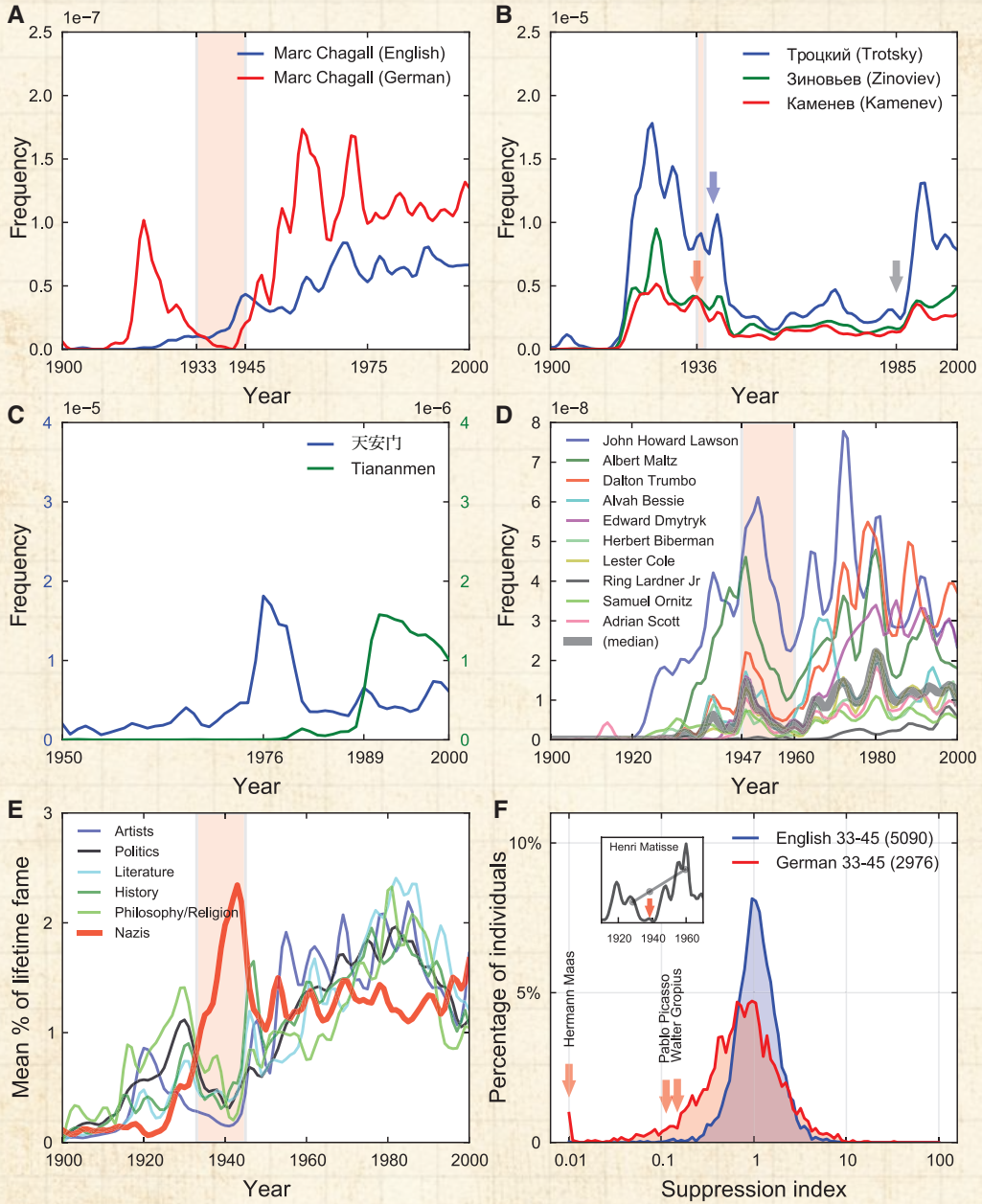
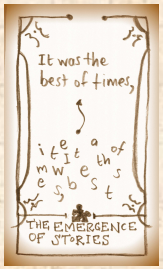
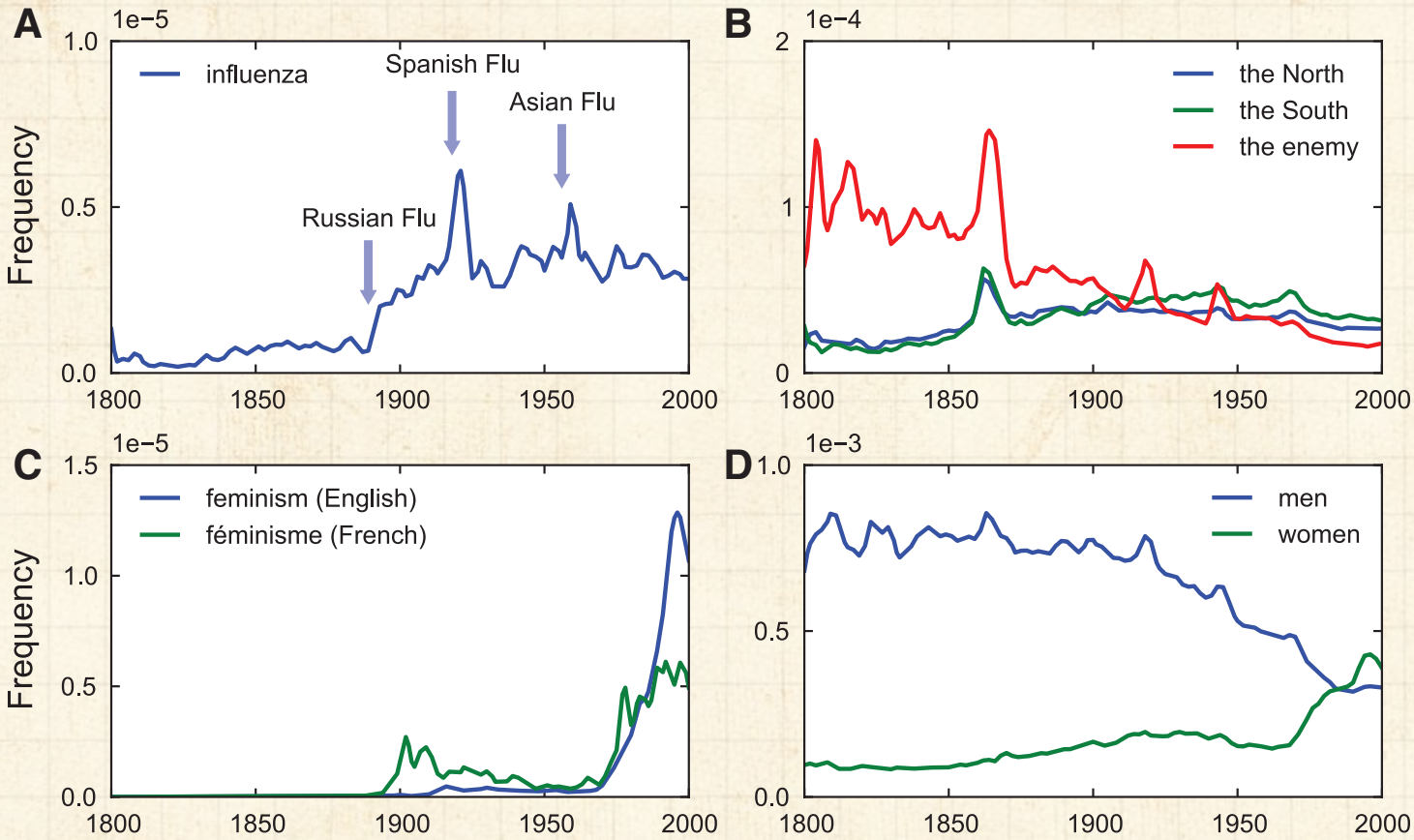


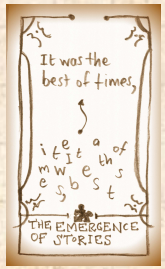
Fig. 4. Culturomics can be used to detect censorship. (**A**) Usage frequency of "Marc Chagall" in German (red) as compared to English (blue). (**B**) Suppression of Leon Trotsky (blue), Grigory Zinoviev (green), and Lev Kamenev (red) in Russian texts, with noteworthy events indicated: Trotsky's assassination (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

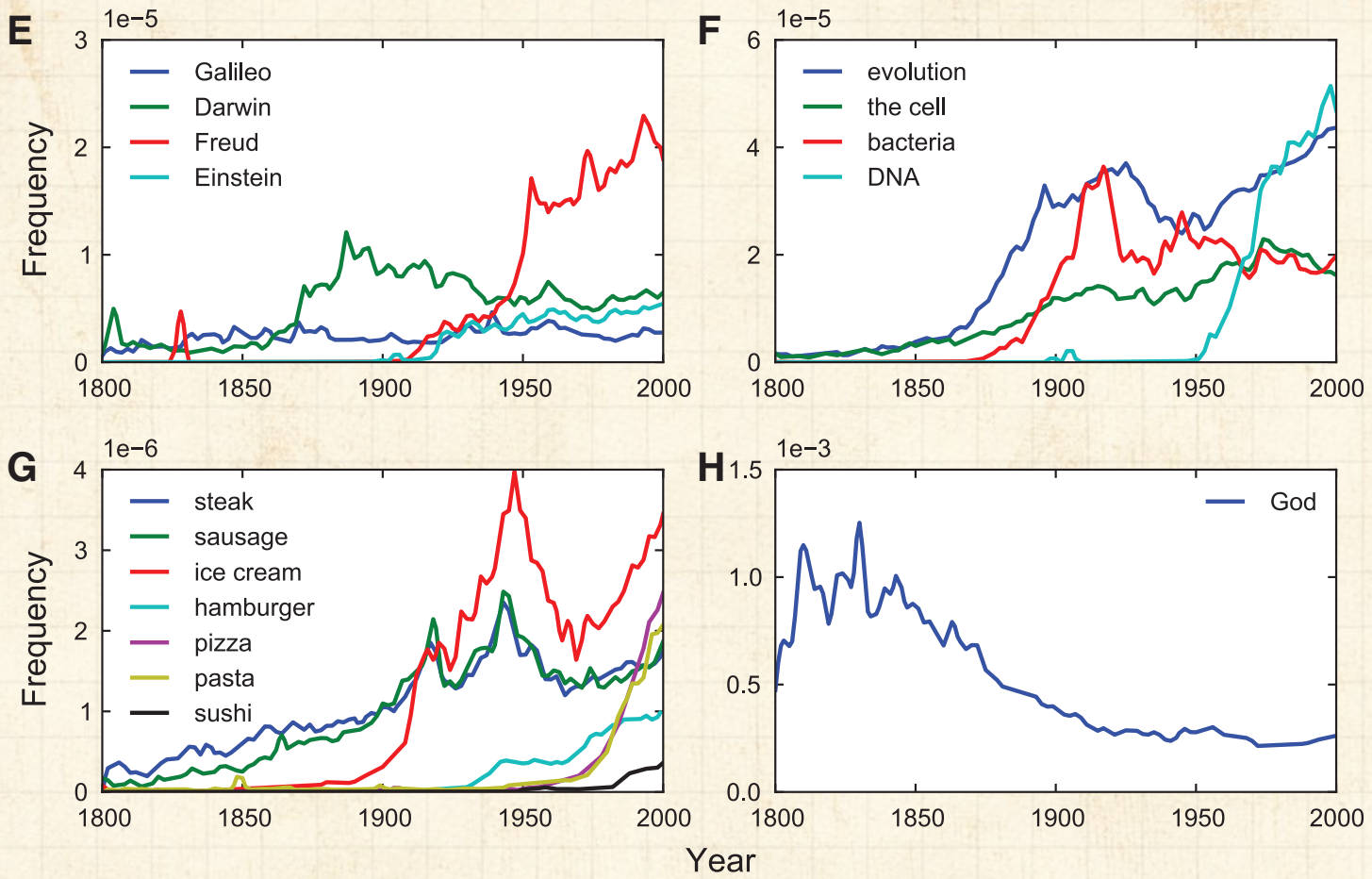
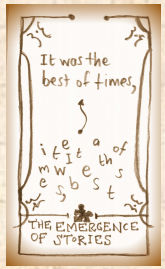
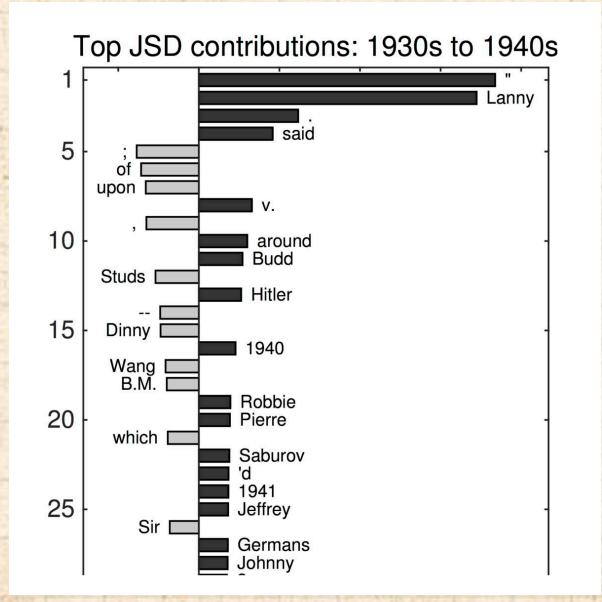


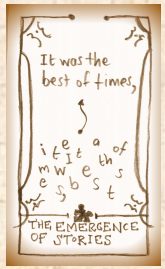
Fig. 5. Culturomics provides quantitative evidence for scholars in many fields. (**A**) Historical epidemiology: "influenza" is shown in blue; the Russian, Spanish, and Asian flu epidemics are highlighted. (**B**) History of the Civil War. (**C**) Comparative history. (**D**) Gender studies. (**E** and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



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

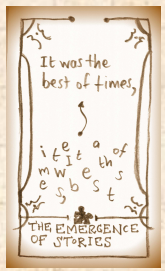
Pechenick, Danforth, and Dodds, PLoS ONE, **10**, e0137041, 2015. [2]



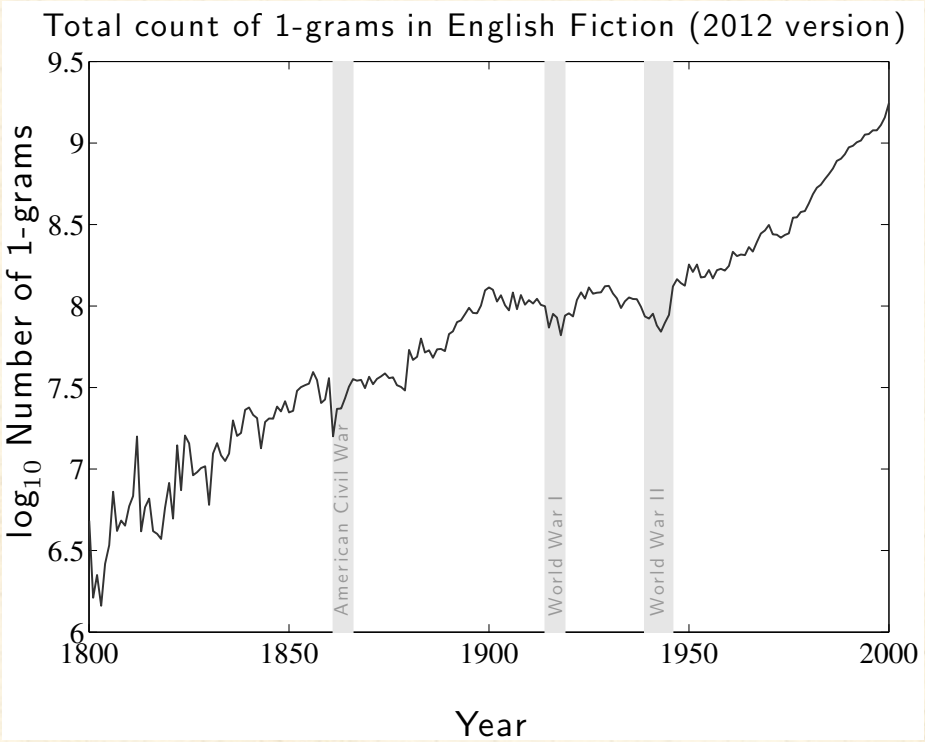


Press:

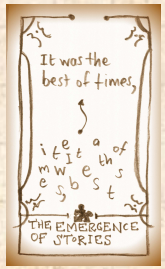
- New York Times: Google Books: A Complex and Controversial Experiment by Stephen Heyman (October 28, 2015)
- Future Tense, slate.com: Is Google Books Leading Researchers Astray? by Jacob Brogan (October 14, 2015)
- wired.com: The pitfalls of using Google Ngram to study language by Sarah Zhang (October 12, 2015)
- discovery.com Can Google Books Really Tell Us About Cultural Evolution? by Neuroskeptic (October 10, 2015)

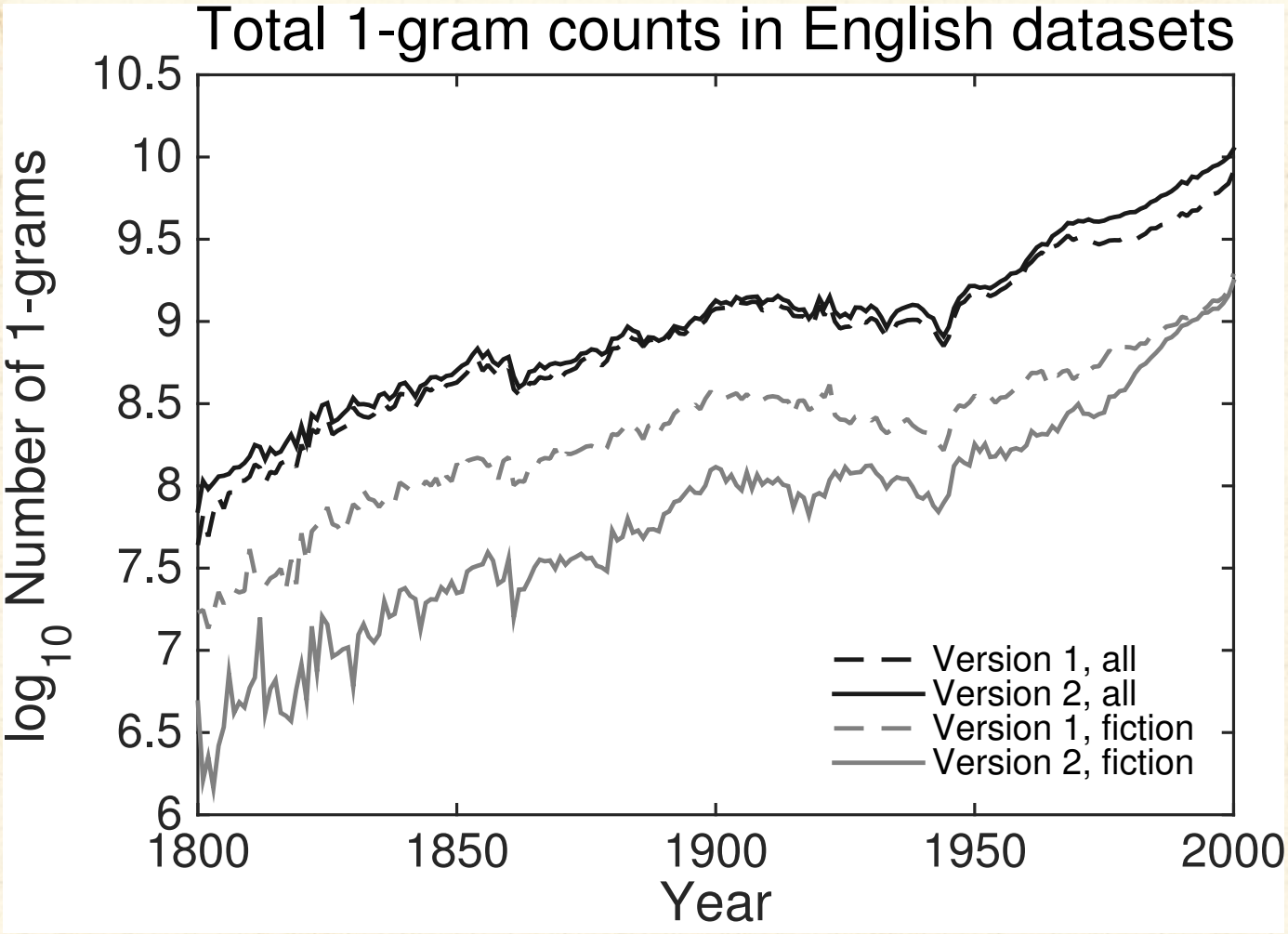


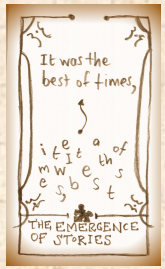
Volume of "words"—exponential growth



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



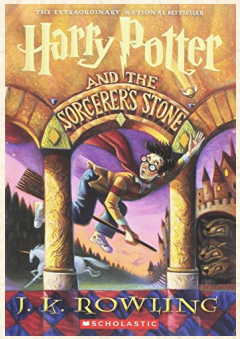

Total 1-gram counts in English datasets
log10 Number of 1-grams
10.5
10
9.5
9
8.5
8
7.5
7
6.5
6
1800
1850
1900
1950
2000
Year
Version 1, all
Version 2, all
Version 1, fiction
Version 2, fiction


It was the best of times,
THE EMERGENCE OF STORIES

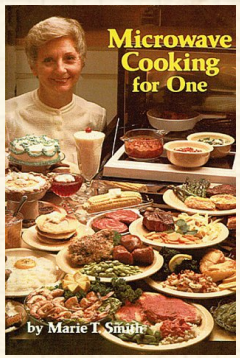
Trouble at Mill, 1/2:

Every book gets one vote:

- Equally important:

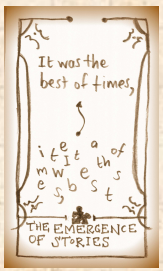


"Harry Potter and the Sorcerer's Stone" by J. K. Rowling (1998). [3]



"Microwave Cooking for One" by Marie Smith (1999). [4]

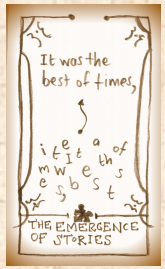
- New editions, revisions, reprintings give very modest bump.



Trouble at Mill, 2/2:

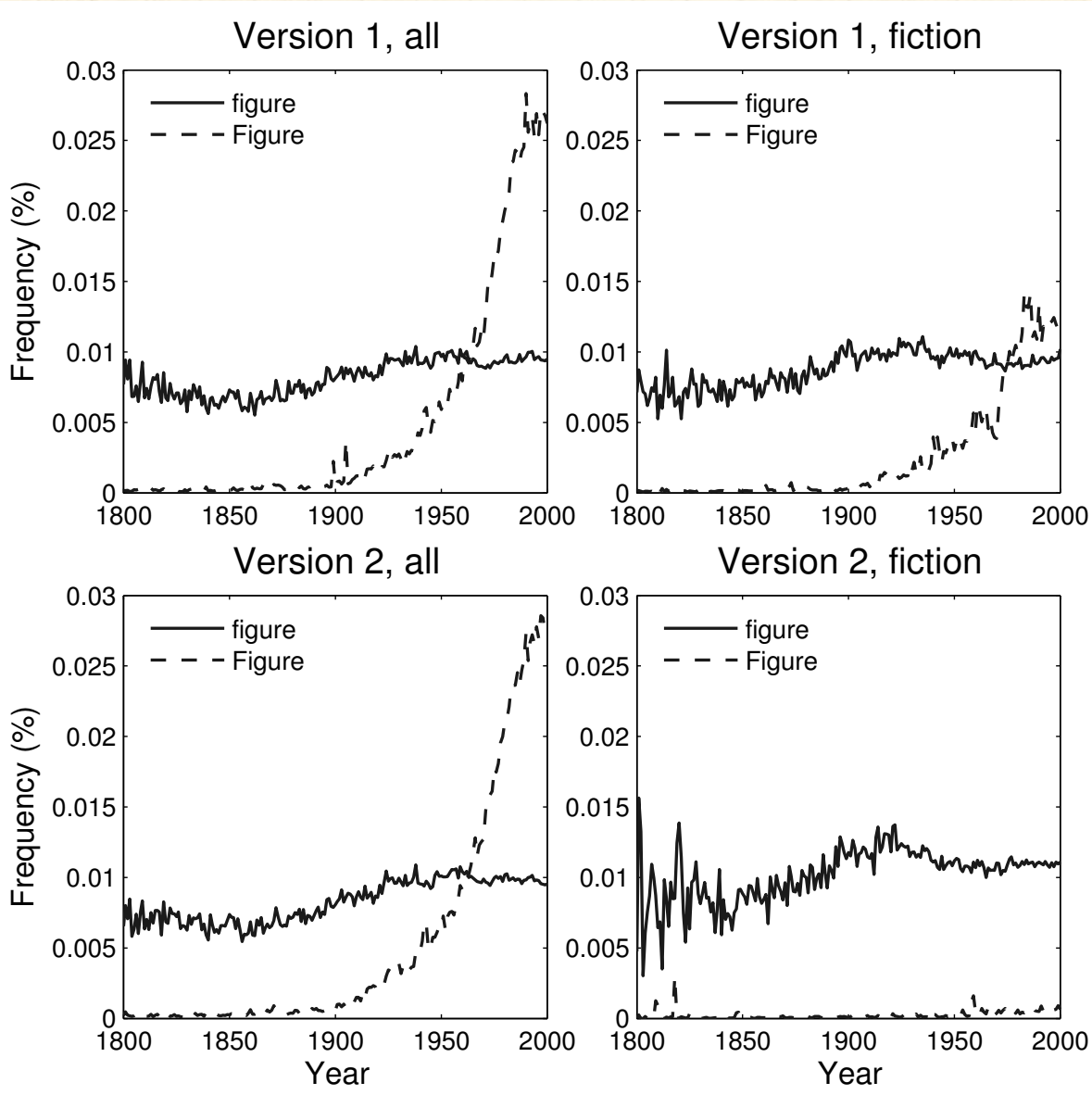
Lord of the Rings is fading away:

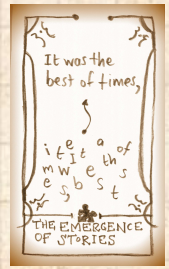
- Search for Frodo, Gandalf in English Fiction, 2012.
- English Fiction = fiction + literary criticism.



Trouble at Mill, 2/2:

Google Books inhaled a lot of Science:



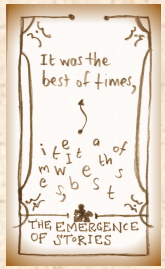


Kullback-Leibler divergence:

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:

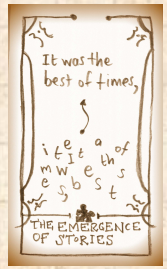
$$D_{\mathsf{JS}}(P\,||\,Q) = \frac{1}{2}\left(D_{KL}(P||M) + D_{KL}(Q||M)\right),$$

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

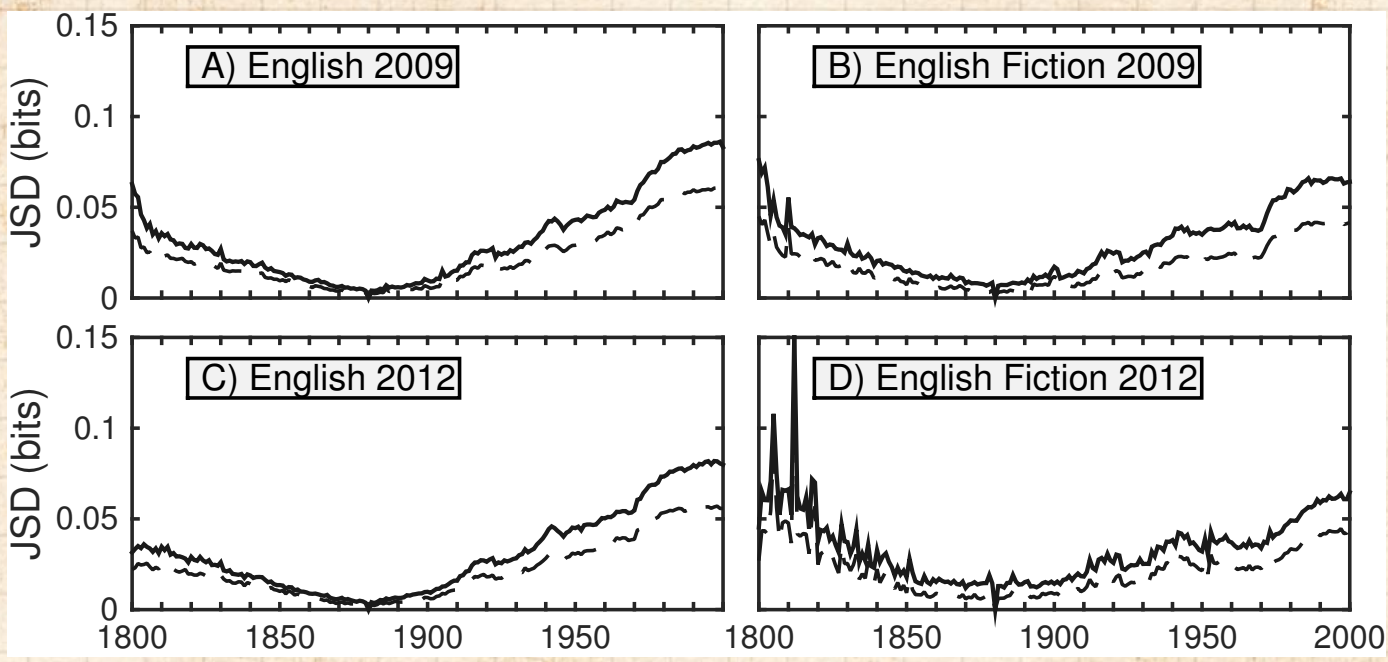
$$D_{\mathsf{JS}}(P\,||\,Q) = H(M) - \frac{1}{2}\left(H(P) + H(Q)\right)$$

- Use per word contribution to the JSD to make shifts:

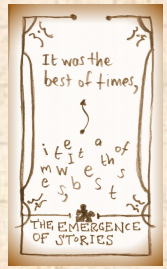
$$D_{\mathsf{JS},i}(P\,||\,Q) = -m_i \log_2 m_i + \frac{1}{2}\left(p_i \log_2 p_i + q_i \log_2 q_i\right)$$



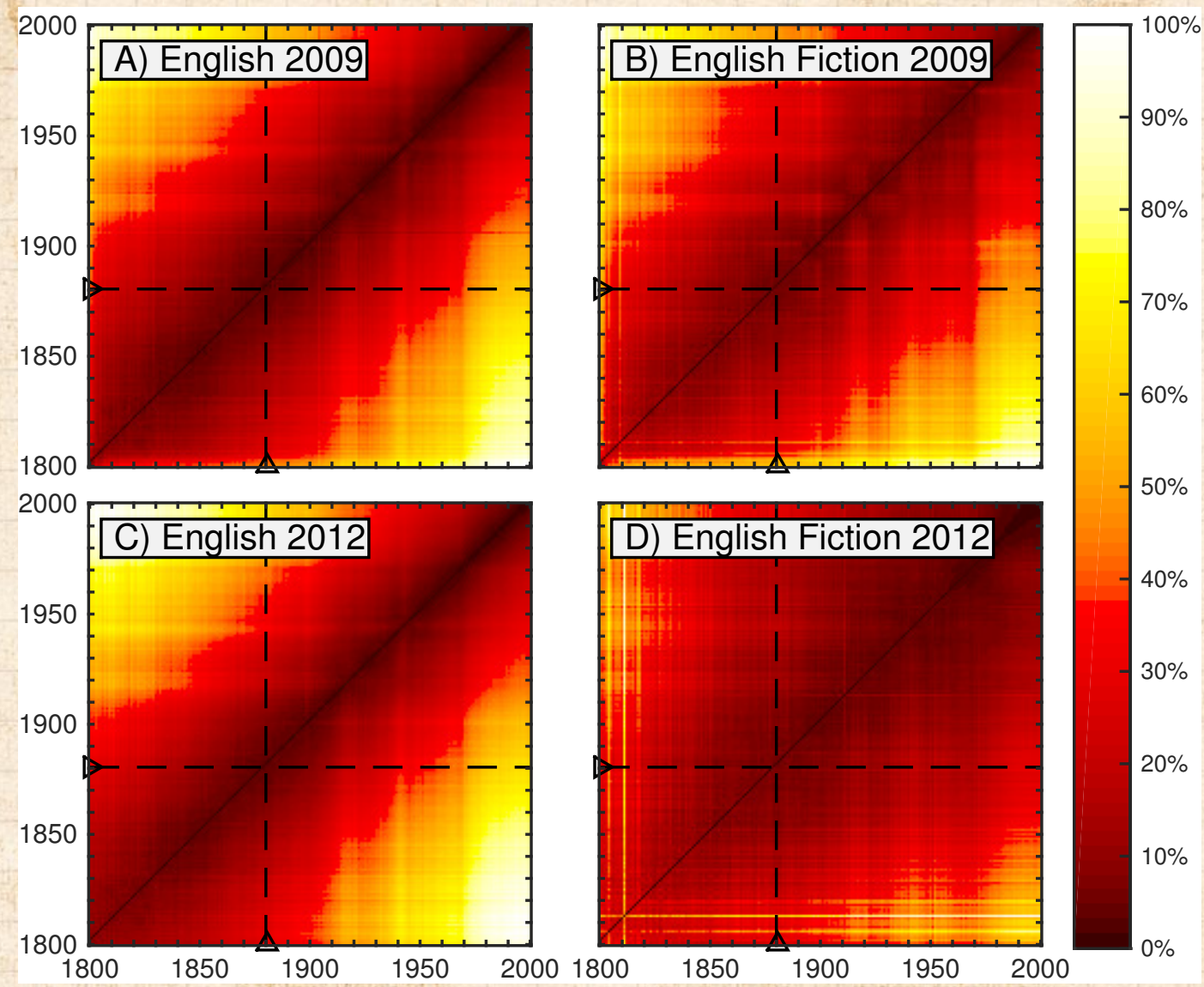
JSD between 1880 and 1800–2000:

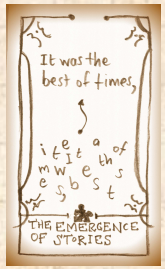


Contributions are counted for all words appearing above a 10^{-5} threshold in a given year; for the dashed curves, the threshold is 10^{-4}.

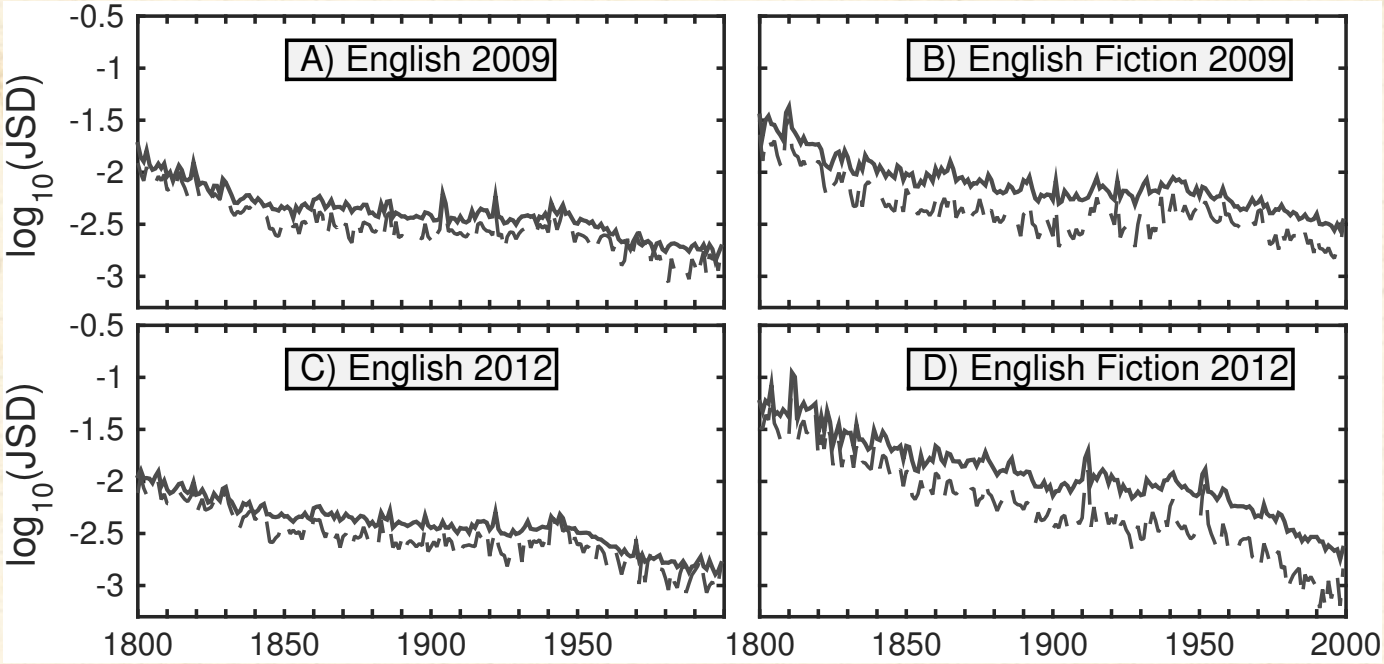


JSD between years:

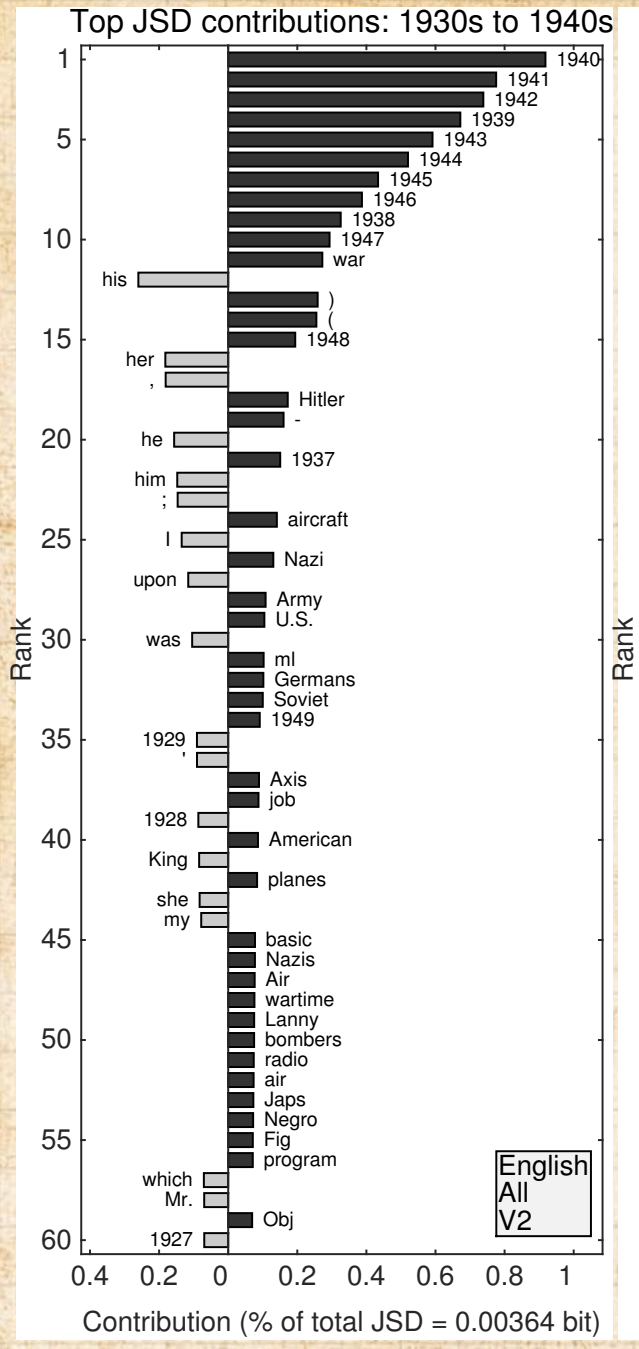


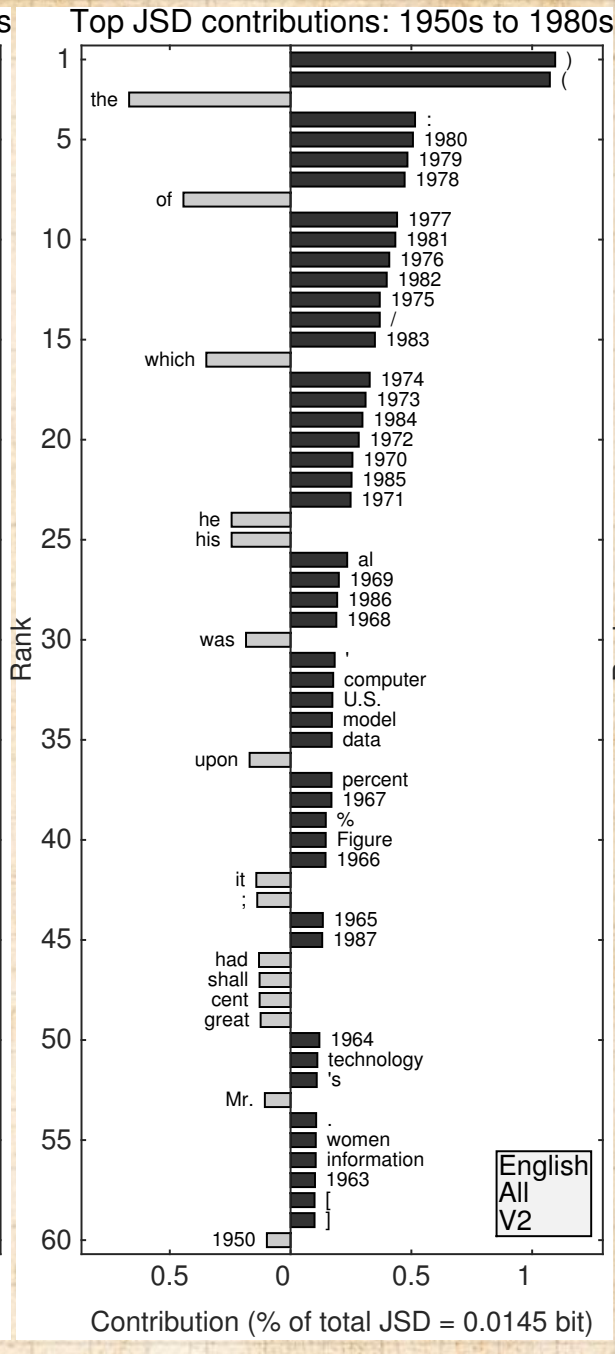


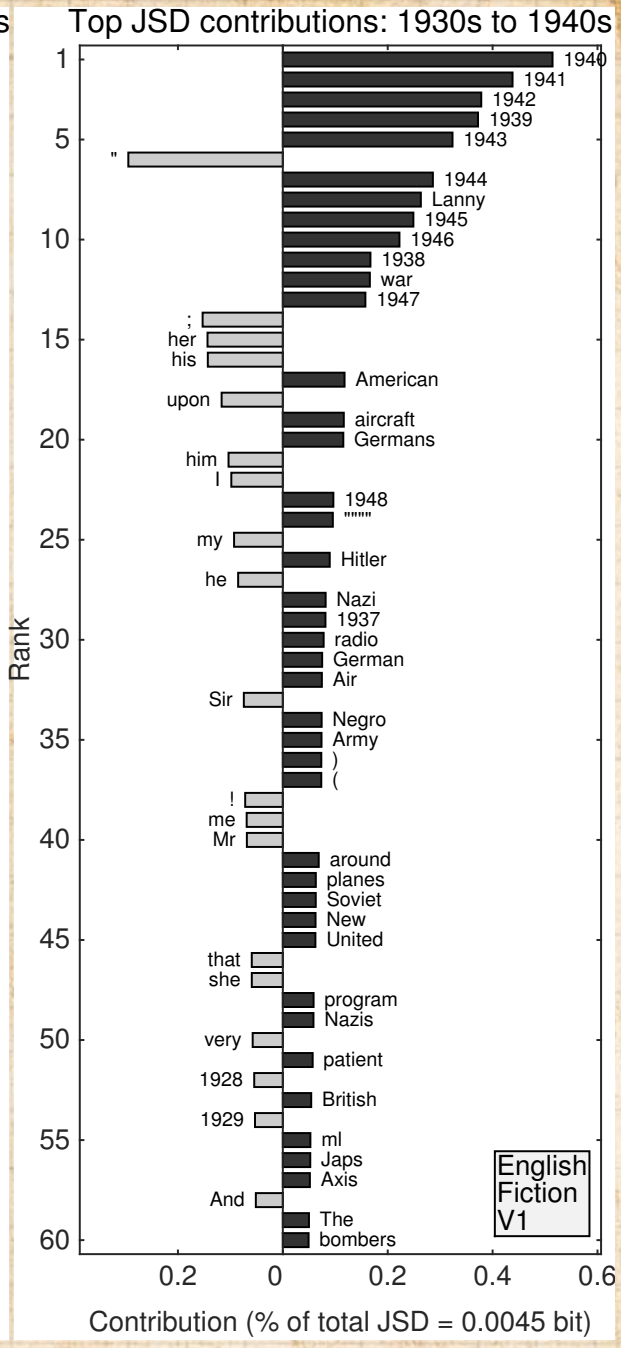
JSD between consecutive years:

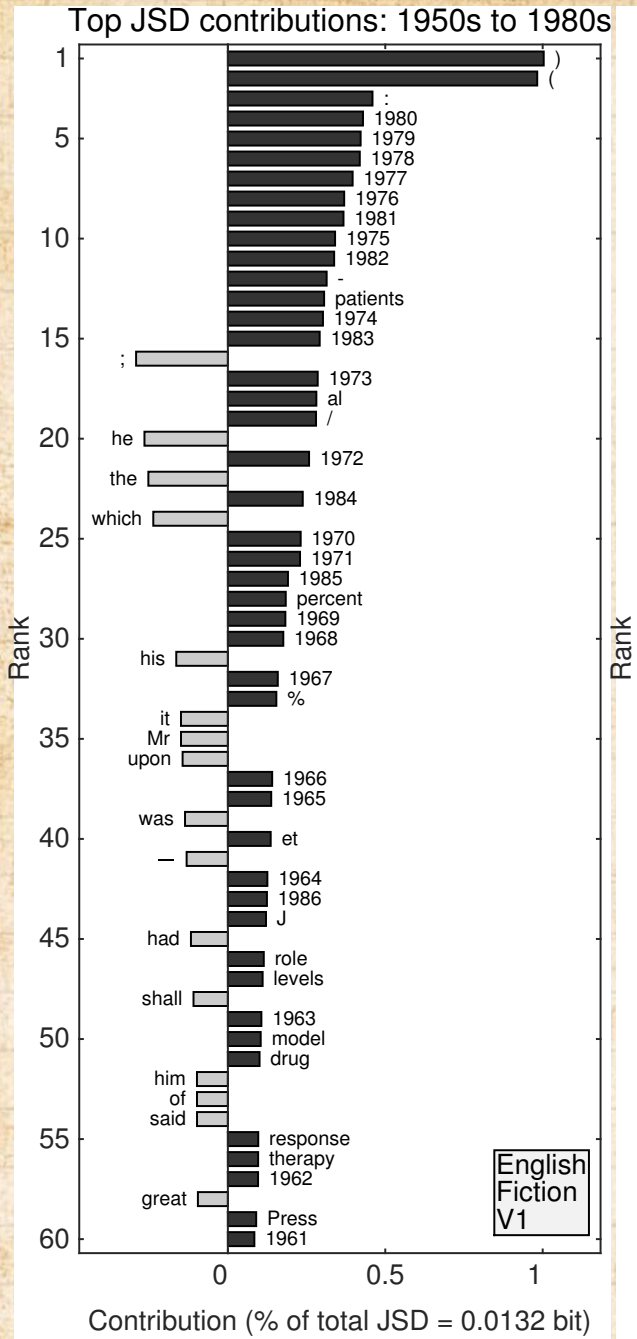


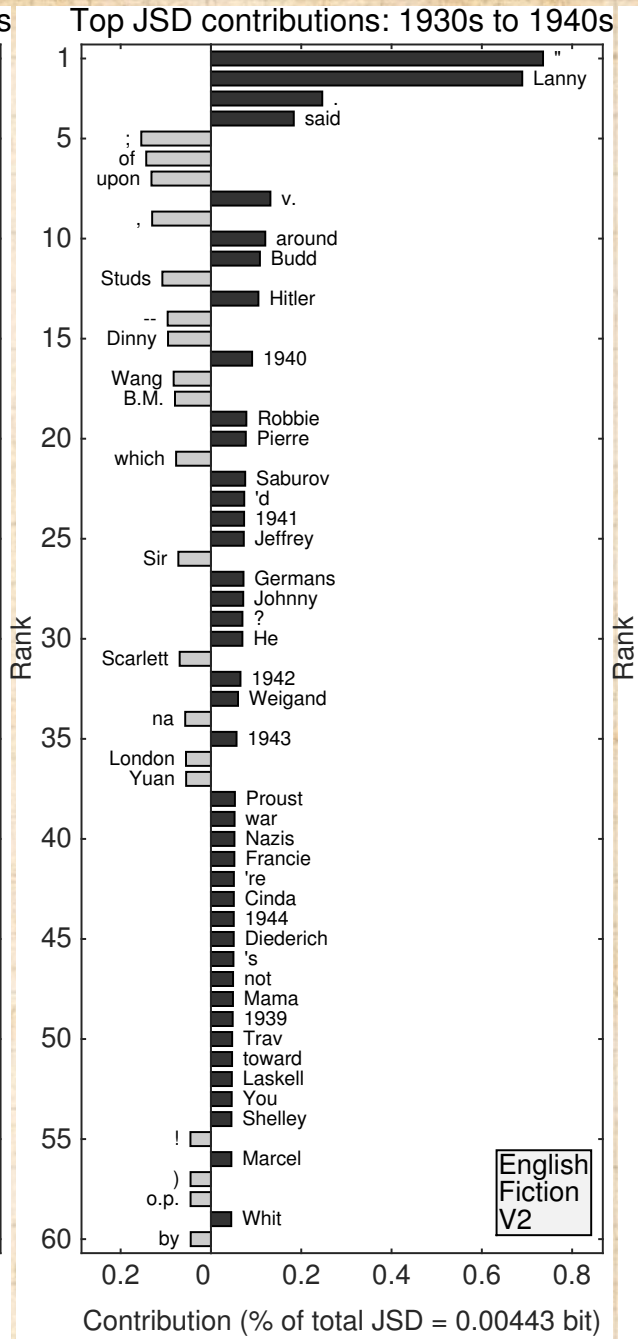
Consecutive year (between each year and the following year) base-10 logarithms of JSD, corresponding to off-diagonals. For the solid curves, contributions are counted for all words appearing above a 10^{-5} threshold in a given year; for the dashed curves, the threshold is 10^{-4}. Divergences between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

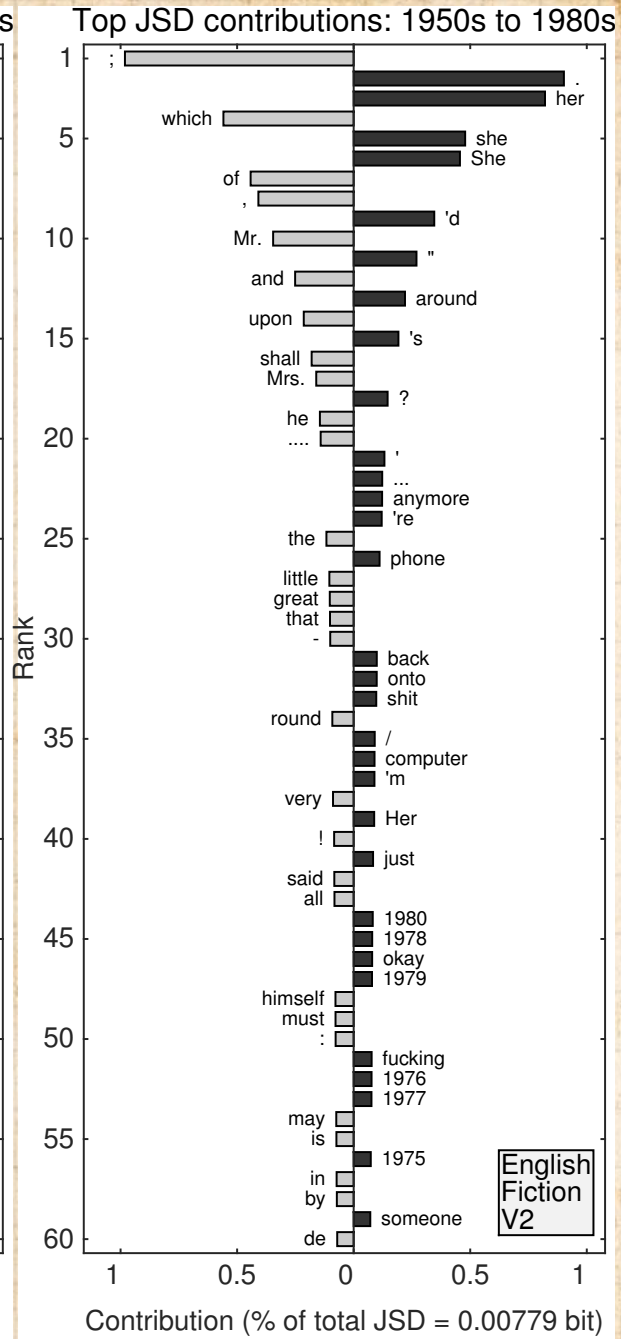

Top JSD contributions: 1930s to 1940s
Rank
1940
1941
1942
1939
1943
1944
1945
1946
1938
1947
war
his
)
(
1948
her
,
Hitler
-
he
1937
him
;
aircraft
I
Nazi
upon
Army
U.S.
was
ml
Germans
Soviet
1949
1929
'
Axis
job
1928
American
King
planes
she
my
basic
Nazis
Air
wartime
Lanny
bombers
radio
air
Japs
Negro
Fig
program
which
Mr.
Obj
1927
English
All
V2
Contribution (% of total JSD = 0.00364 bit)


Top JSD contributions: 1950s to 1980s
Rank
)
(
the
:
1980
1979
1978
of
1977
1981
1976
1982
1975
/
1983
which
1974
1973
1984
1972
1970
1985
1971
he
his
al
1969
1986
1968
was
'
computer
U.S.
model
data
upon
percent
1967
%
Figure
1966
it
;
1965
1987
had
shall
cent
great
1964
technology
's
Mr.
.
women
information
1963
[
]
1950
English
All
V2
Contribution (% of total JSD = 0.0145 bit)

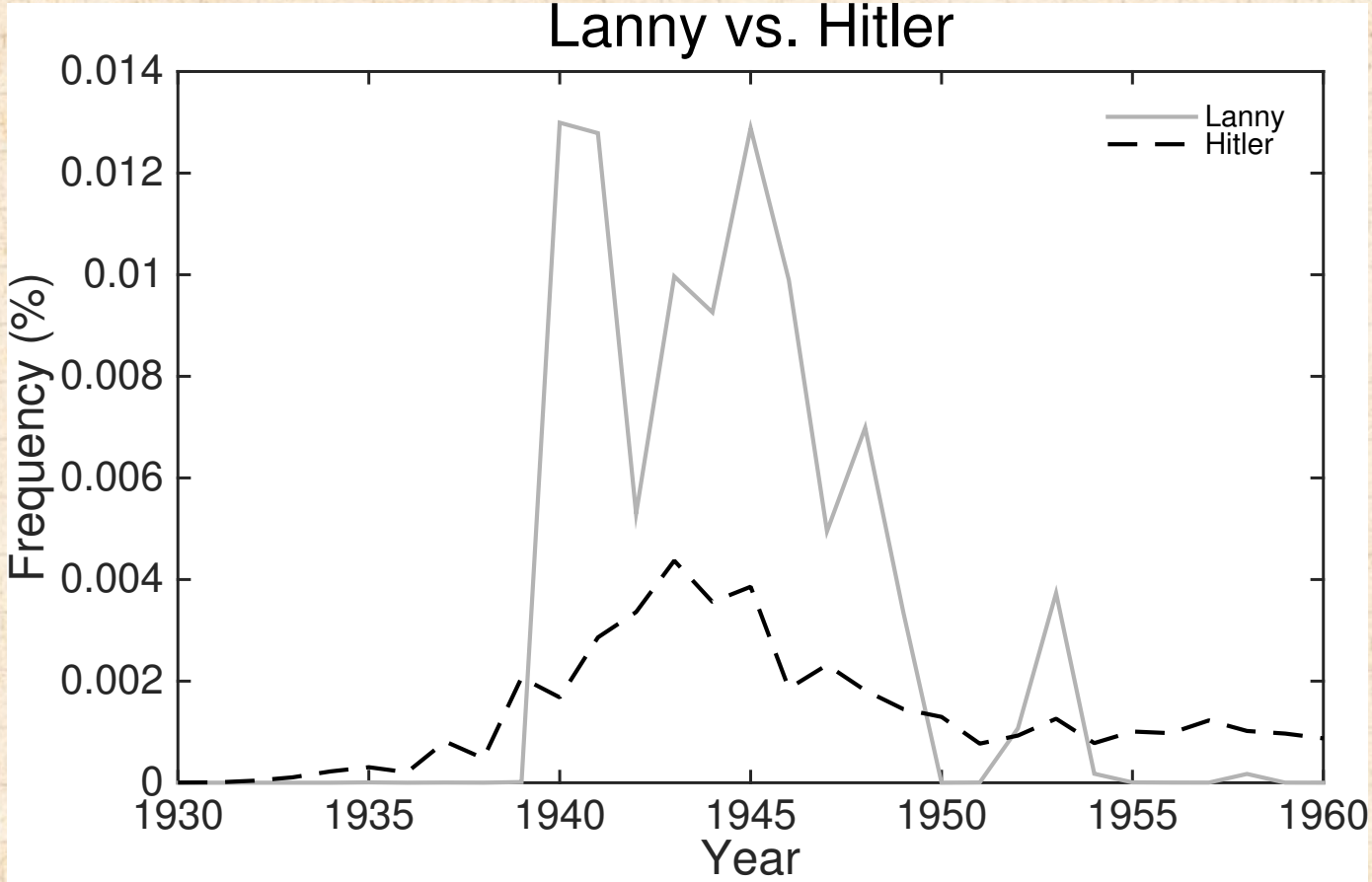

Top JSD contributions: 1930s to 1940s
Rank
1940
1941
1942
1939
1943
"
1944
Lanny
1945
1946
1938
war
1947
;
her
his
American
upon
aircraft
Germans
him
I
1948
"""
my
Hitler
he
Nazi
1937
radio
German
Air
Sir
Negro
Army
)
(
!
me
Mr
around
planes
Soviet
New
United
that
she
program
Nazis
very
patient
1928
British
1929
ml
Japs
Axis
And
The
bombers
English
Fiction
V1
Contribution (% of total JSD = 0.0045 bit)

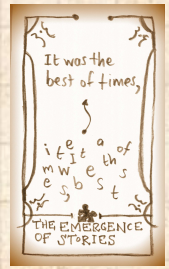

Top JSD contributions: 1950s to 1980s
Rank
Contribution (% of total JSD = 0.0132 bit)
English Fiction V1
)
(
:
1980
1979
1978
1977
1976
1981
1975
1982
-
patients
1974
1983
;
1973
al
/
he
1972
the
1984
which
1970
1971
1985
percent
1969
1968
his
1967
%
it
Mr
upon
1966
1965
was
et
—
1964
1986
J
had
role
levels
shall
1963
model
drug
him
of
said
response
therapy
1962
great
Press
1961


Top JSD contributions: 1930s to 1940s
Rank
Contribution (% of total JSD = 0.00443 bit)
English Fiction V2
"
Lanny
.
said
;
of
upon
v.
,
around
Budd
Studs
Hitler
--
Dinny
1940
Wang
B.M.
Robbie
Pierre
which
Saburov
'd
1941
Jeffrey
Sir
Germans
Johnny
?
He
Scarlett
1942
Weigand
na
1943
London
Yuan
Proust
war
Nazis
Francie
're
Cinda
1944
Diederich
's
not
Mama
1939
Trav
toward
Laskell
You
Shelley
!
Marcel
)
o.p.
Whit
by

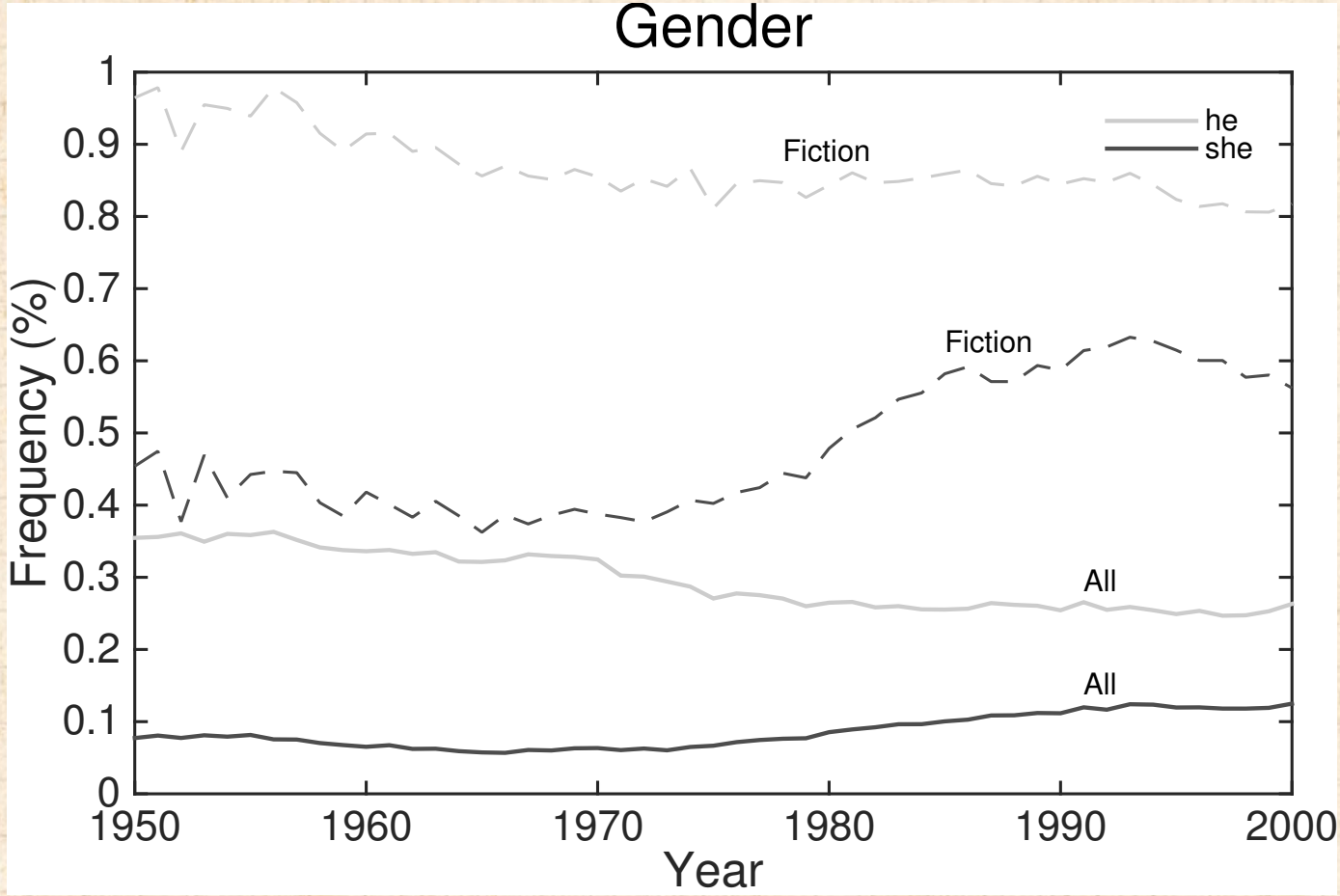

Top JSD contributions: 1950s to 1980s
Rank
Contribution (% of total JSD = 0.00779 bit)
English Fiction V2
;
.
her
which
she
She
of
,
'd
Mr.
"
and
around
upon
's
shall
Mrs.
?
he
....
'
...
anymore
're
the
phone
little
great
that
-
back
onto
shit
round
/
computer
'm
very
Her
!
just
said
all
1980
1978
okay
1979
himself
must
:
fucking
1976
1977
may
is
1975
in
by
someone
de

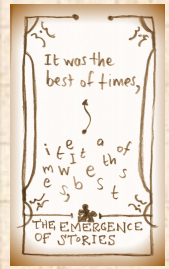
Lanny Budd, Upton Sinclair's forgotten hero



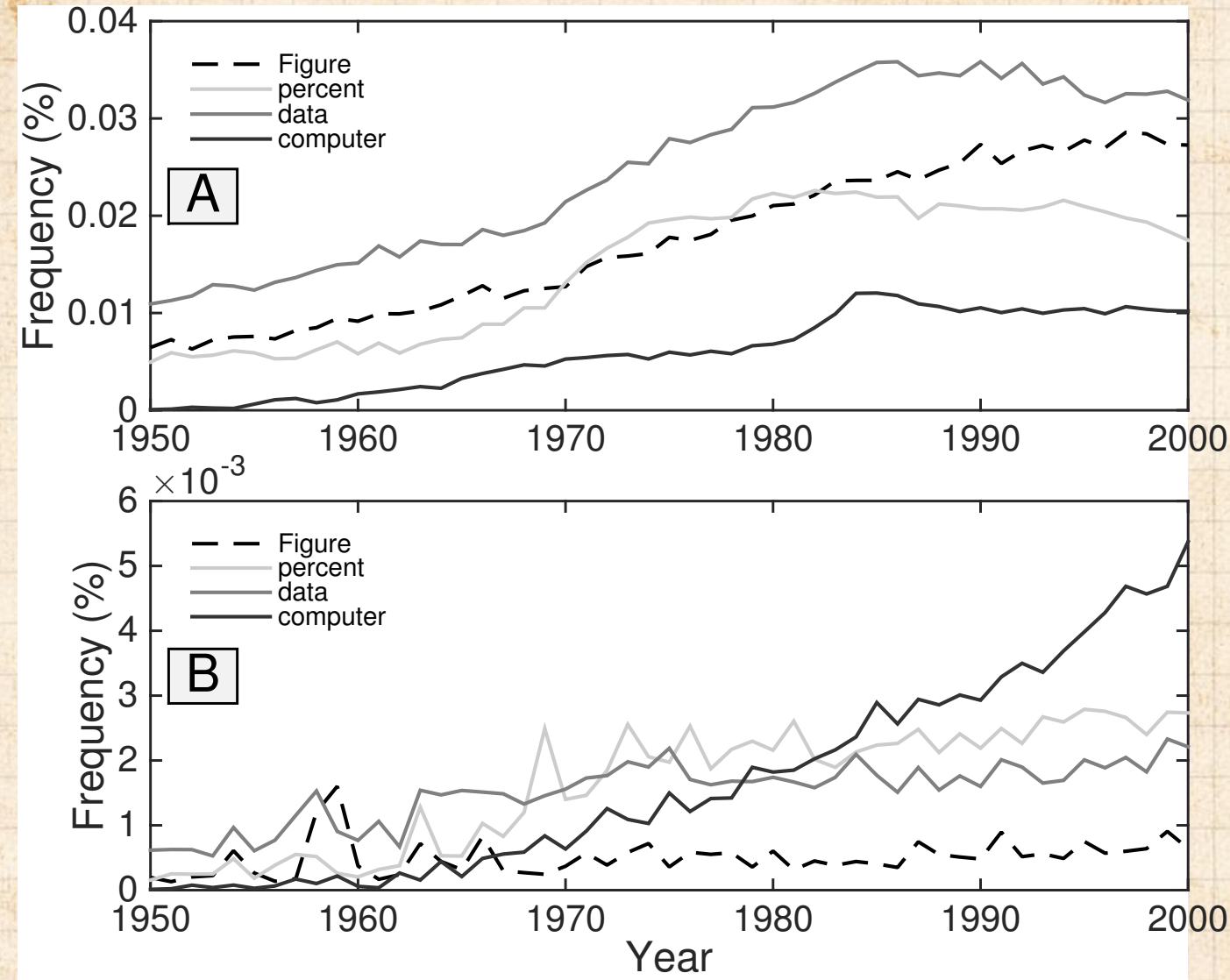


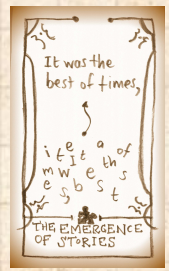
Representative of a more general shift:



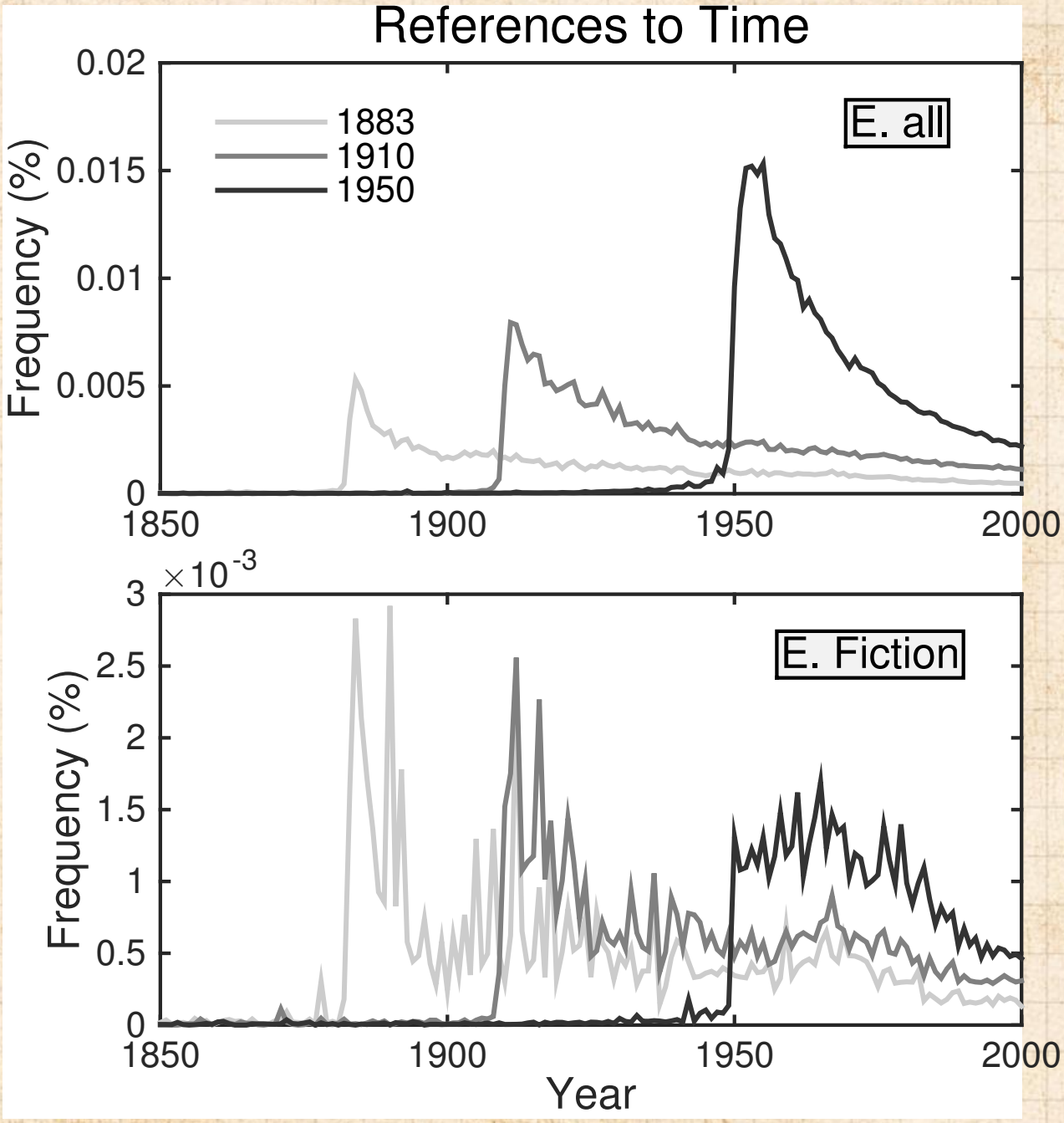


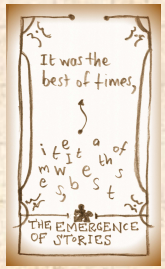
More Science:





Science drives the memory story:





“God is dying”—Google Books

Opinion

It’s Getting Harder to Talk About God

The decline in our spiritual vocabulary has many real-world consequences.

By Jonathan Merritt
Mr. Merritt writes about the intersection of religion, culture and politics.

Oct. 13, 2018

Jeff Rogers

More than 70 percent of Americans identify as Christian, but you wouldn’t know it from listening to them. An overwhelming majority of people say that they don’t feel comfortable speaking about faith, most of the time.

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-sprituality-christian.html

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

Wikipedia's entry on Google ngrams:

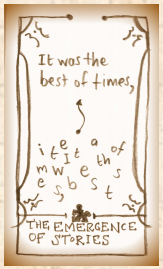
Criticism [edit]

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.[12][13] Because of these errors, and because it is uncontrolled for bias[14] (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.[15] Since the data set does not include metadata, it may not reflect general linguistic or cultural change[16] 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.[14]

OCR issues [edit]

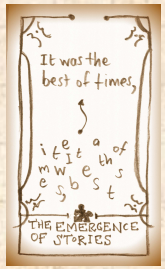
Optical character recognition, or OCR, is not always reliable, and some characters may not be scanned correctly. In particular, systemic errors like the confusion of "s" and "f" in pre-19th century texts (due to the use of the long_s which was similar in appearance to "f") can cause systemic bias. Although 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.[17][18]



Ref. 14 = Pechenick *et al.* [2]

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.

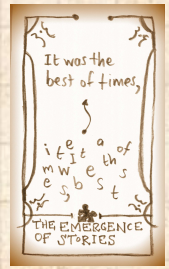


References I

[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

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