Quantitative analysis of culture using millions digitized books

Oleksandra Panova

Taras Shevchenko National University of Kyiv

October 7, 2019

n-gram

n-gram — a contiguous sequence of n items from a given sample of text or speech.
Possible items:

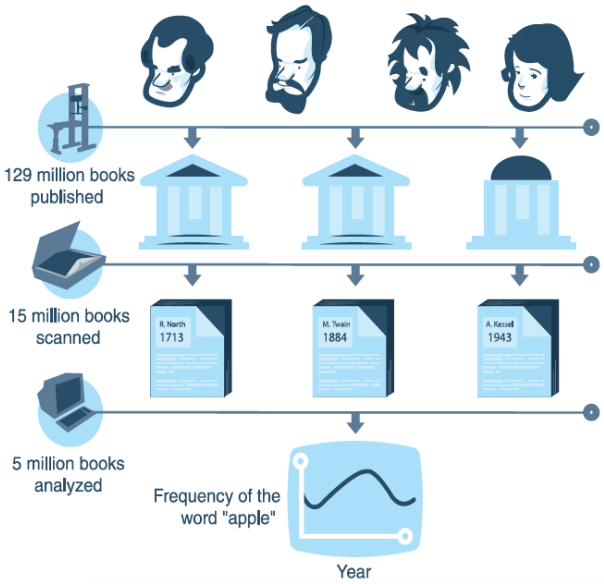
- phonemes,
- letters,
- syllables,
- words.

Classification:

- “unigram” — n-gram of size 1 (“Einstein”),
- “bigram” (“digram”) — n-gram of size 2 (“Albert Einstein”),
- “trigram” — n-gram of size 3 (“physicist Albert Einstein”),
- “four-gram”, “five-gram” and so on — n-gram of size 4, 5 ...

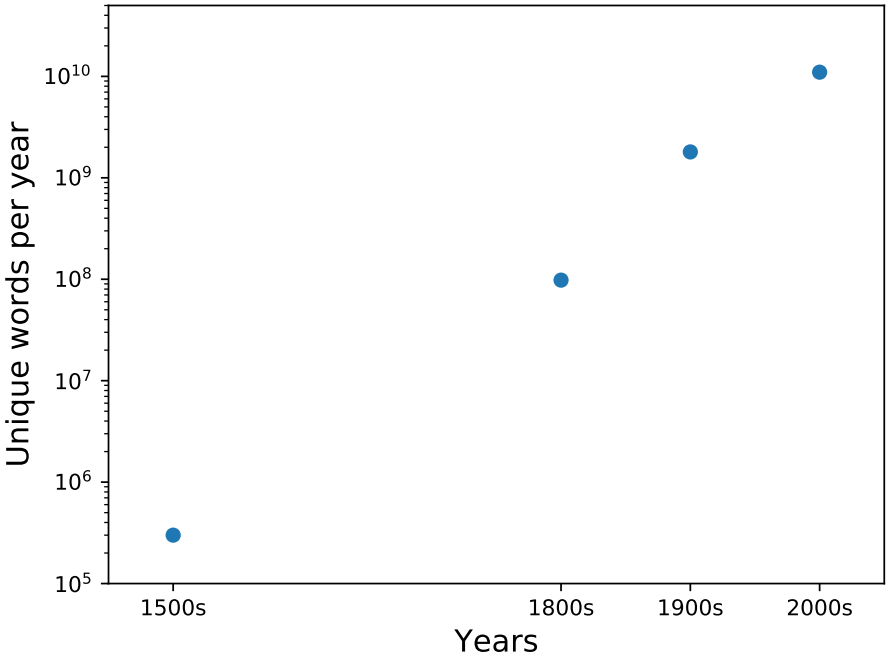
Books digitization

- over 15 million books (∼12% of all books ever printed) have been digitized,
- most books were drawn from over 40 university libraries,
- each page was scanned and digitized by means of optical character recognition (OCR),
- selected 5,195,769 digitized books (4 % of all books ever printed),
- periodicals were excluded.

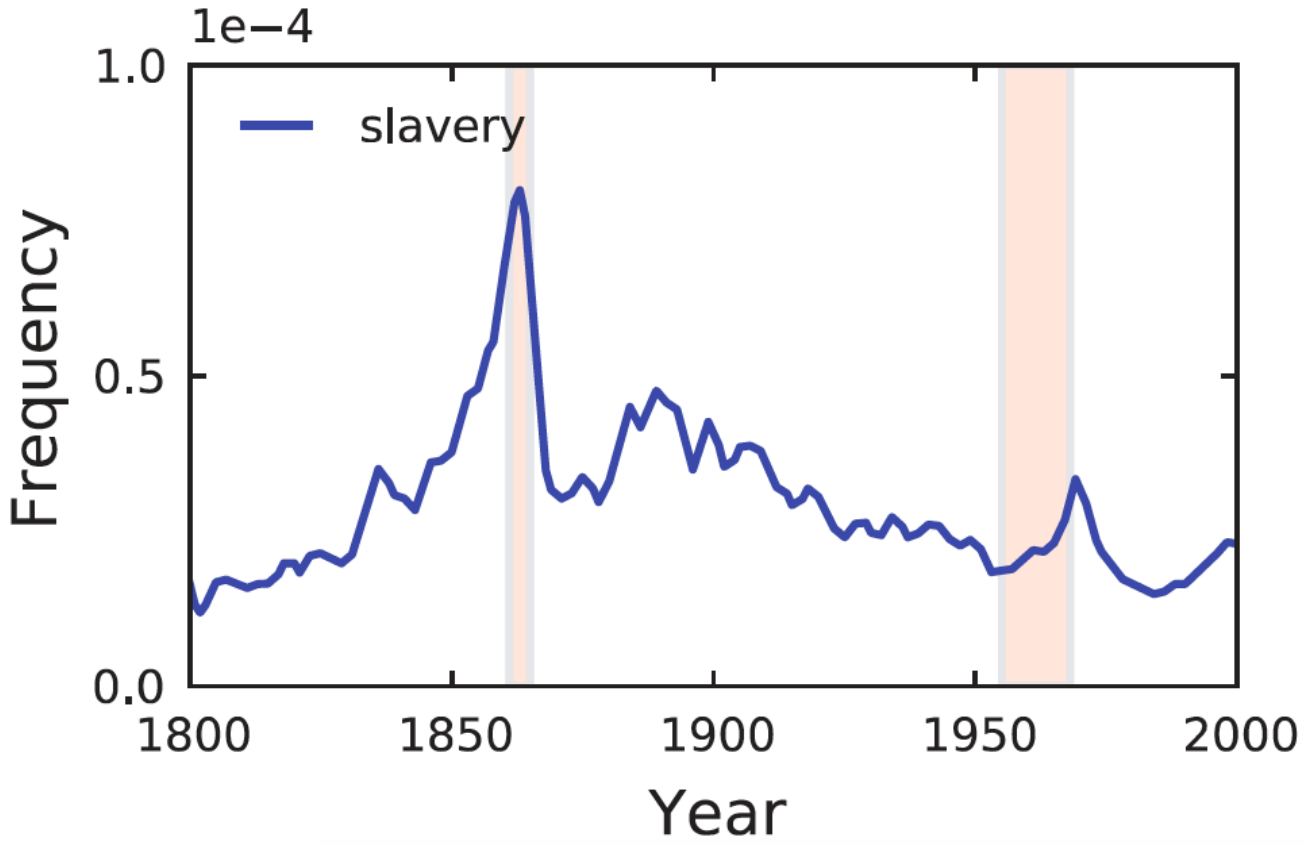


Resulting corpus — 500 billion words

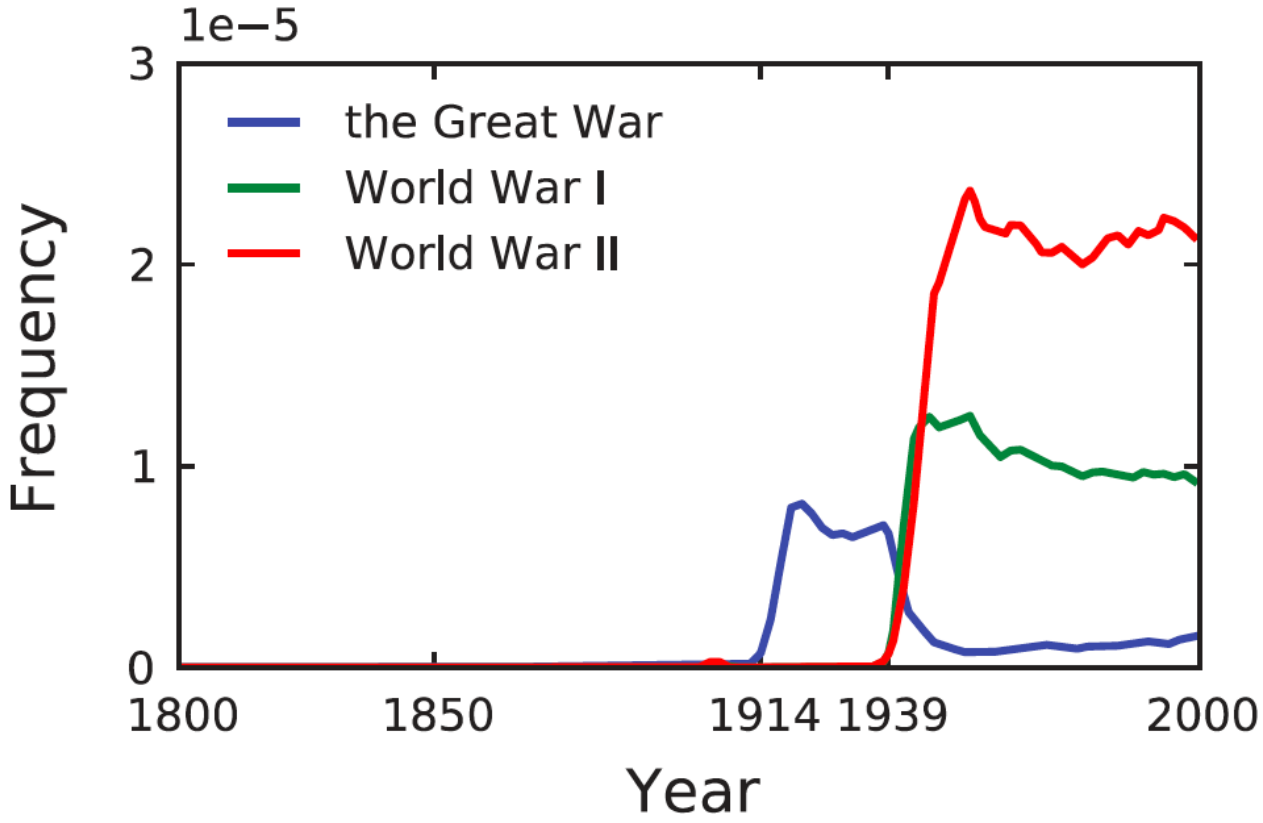
- 361 billion — English,
- 45 billion — French,
- 45 billion — Spanish,
- 37 billion — German,
- 35 billion — Russian,
- 13 billion — Chinese,
- 2 billion — Hebrew.



Cultural change



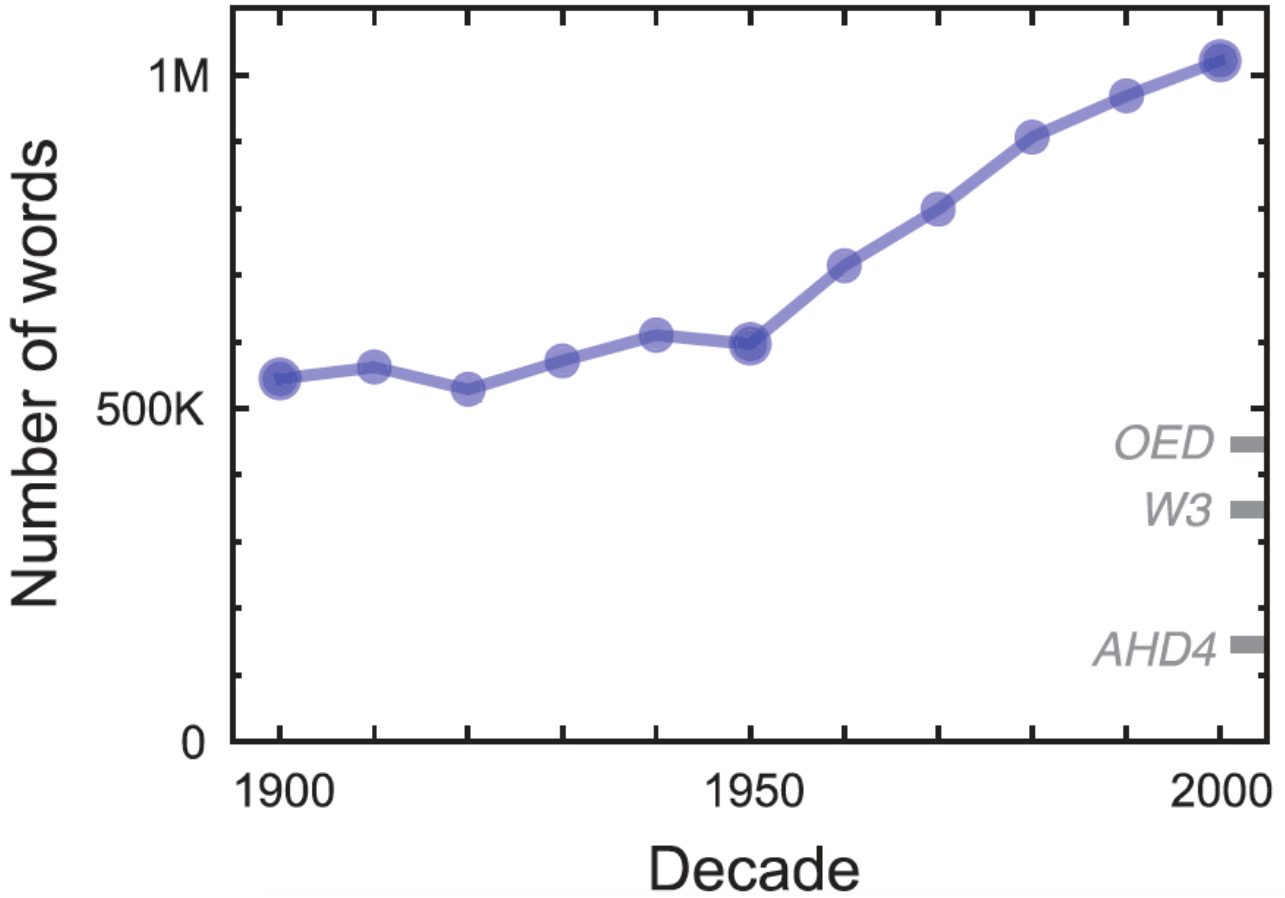
Linguistic change (with cultural roots)



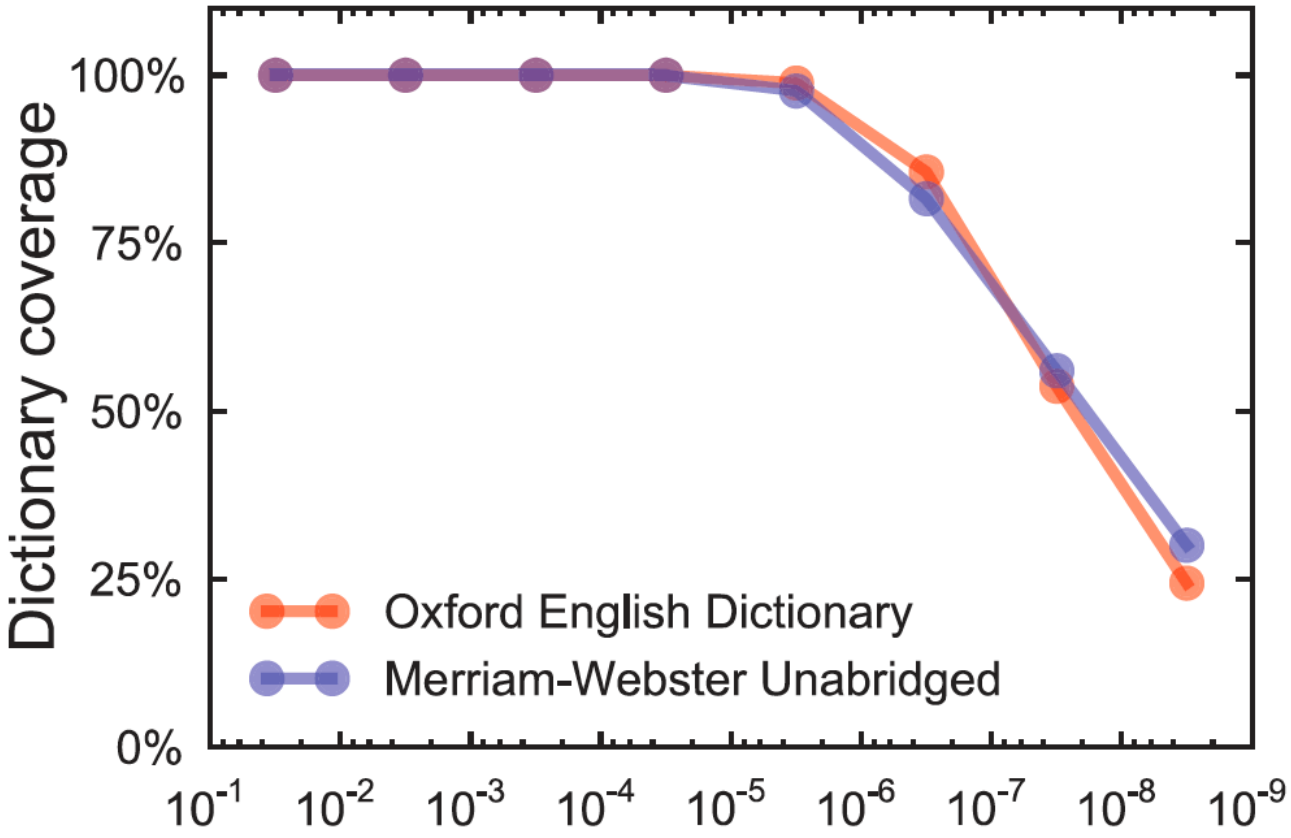
The size of the English lexicon

- 1-gram is common if it's frequency $> 10^{-9}$,
- 1,117,997 common 1-grams in 1900 and 1,489,337 in 2000,
- not all common 1-grams are words,
- manually count the fraction of non-words in random samples: 51% in 1900 and 31% in 2000,
- 544,000 in 1900 and 1,022,000 in 2000,
- increase of lexicon — over 70% during past 50 years (∼8500 words/year),
- we found more words than appear in any dictionary.

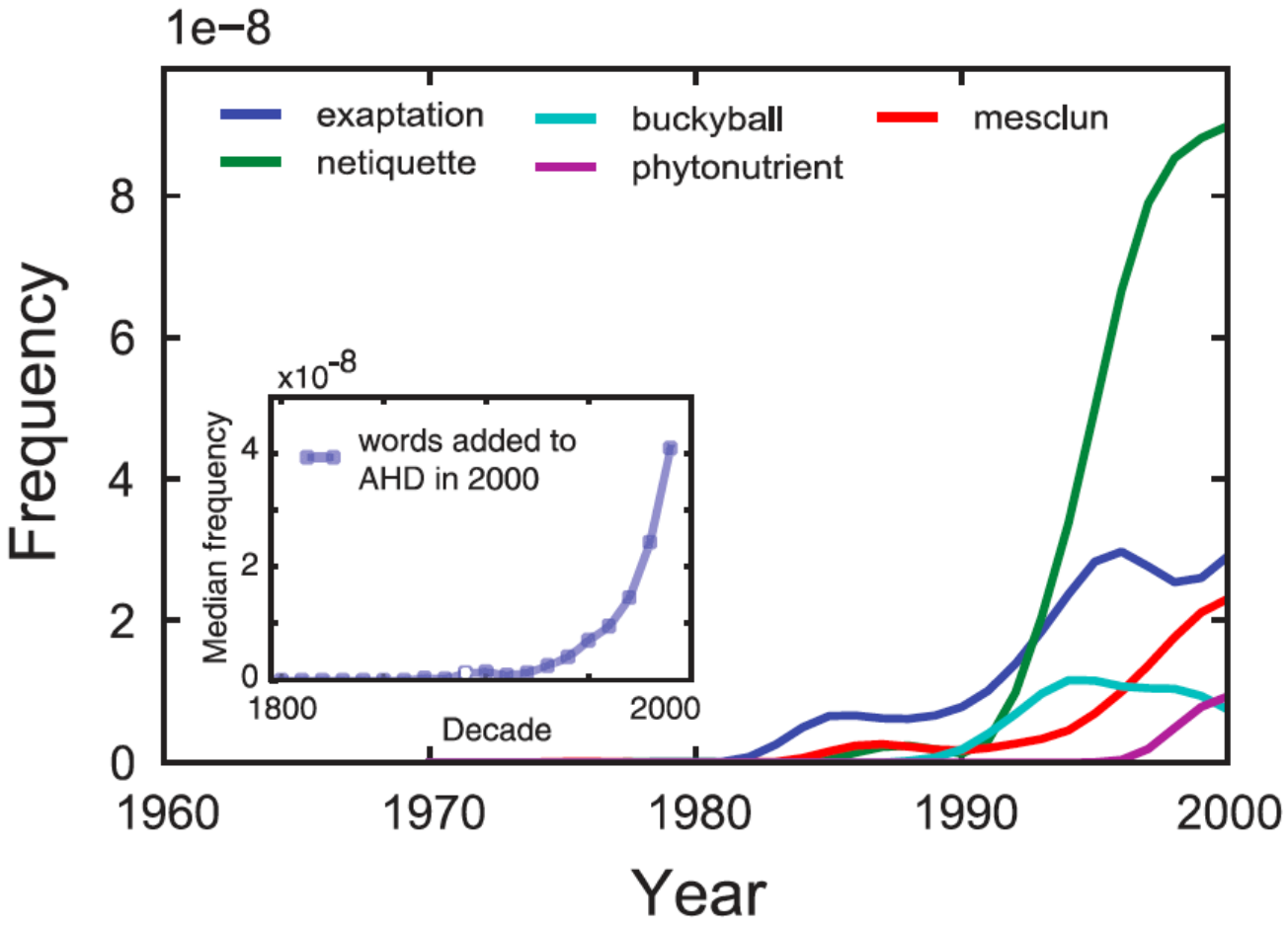
The size of the English lexicon



Dictionary coverage



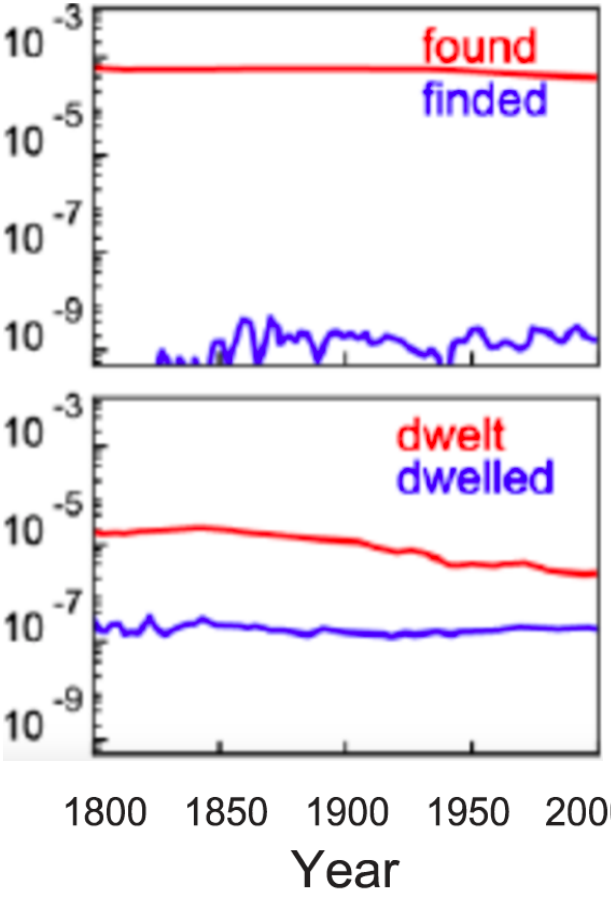
New words

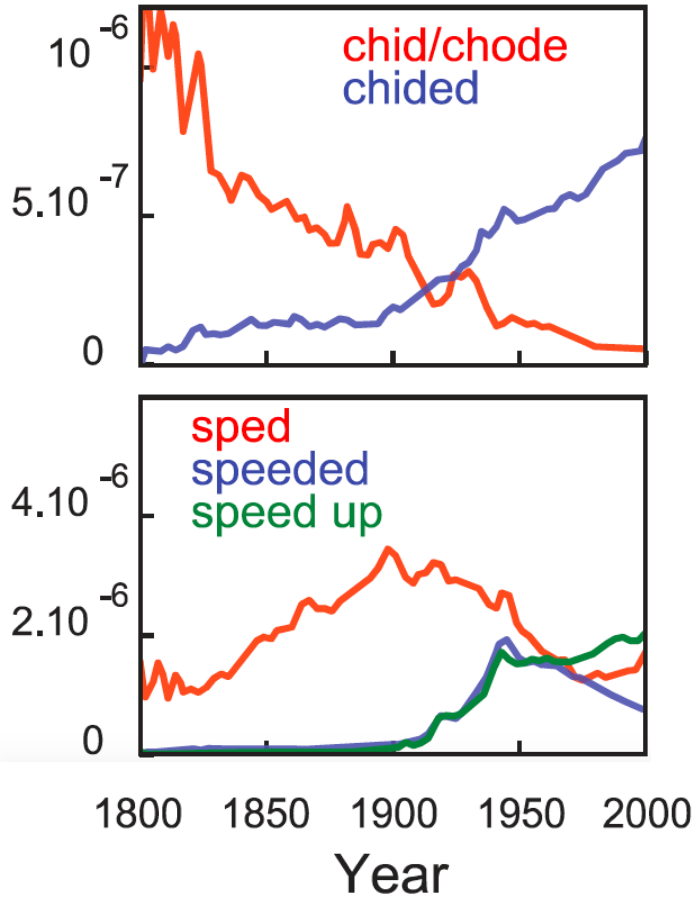

1e−8
exaptation
netiquette
buckyball
phytonutrient
mesclun
Frequency
0
2
4
6
8
1960
1970
1980
1990
2000
Year
x10-8
Median frequency
0
2
4
words added to
AHD in 2000
1800
2000
Decade

How culturomic tools can help lexicographers?

- find low-frequency words in dictionaries and exclude it,
- find new frequent undocumented words and add them to dictionaries,
- provide accurate estimates of current frequency trends to reduce the lag between changes in the lexicon and changes in the dictionaries.

The evolution of grammar: irregular verbs

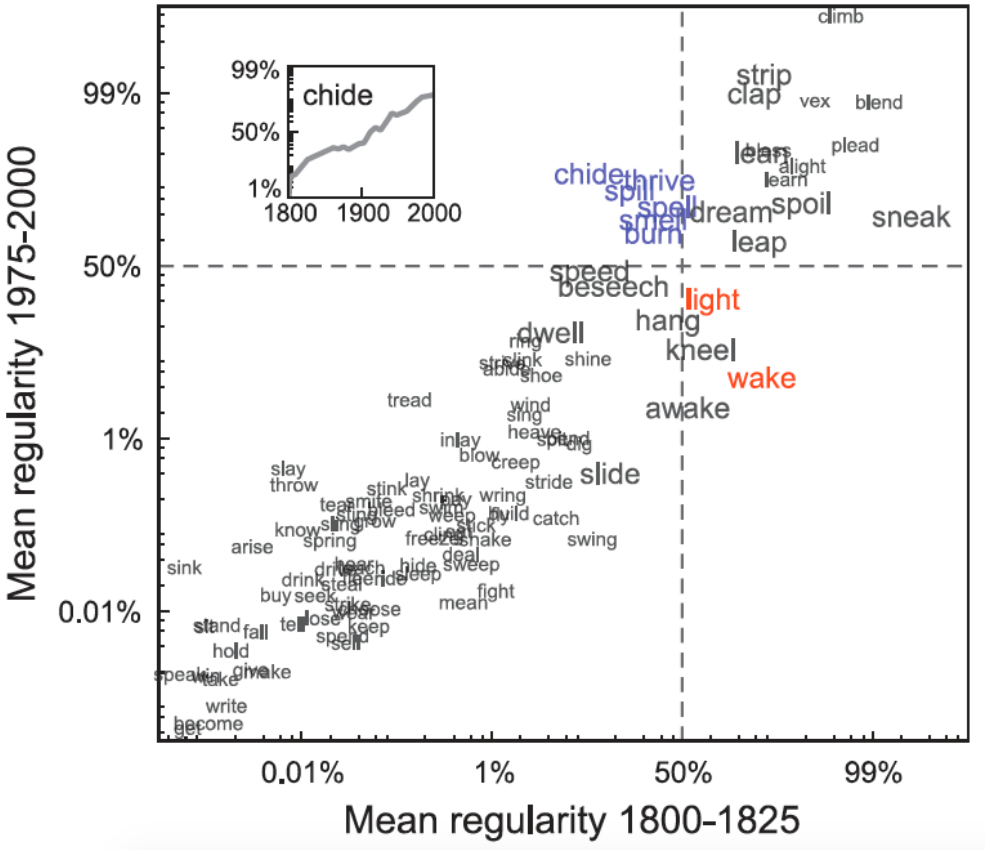




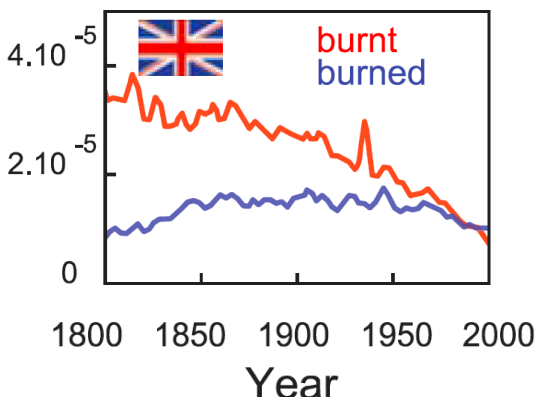
The evolution of grammar: irregular verbs. Regularization

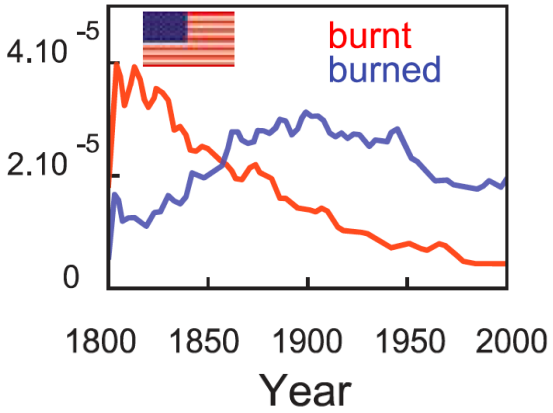
Regularity — percentage of instances in the past tense in which the regular form is used. Regularized verbs:

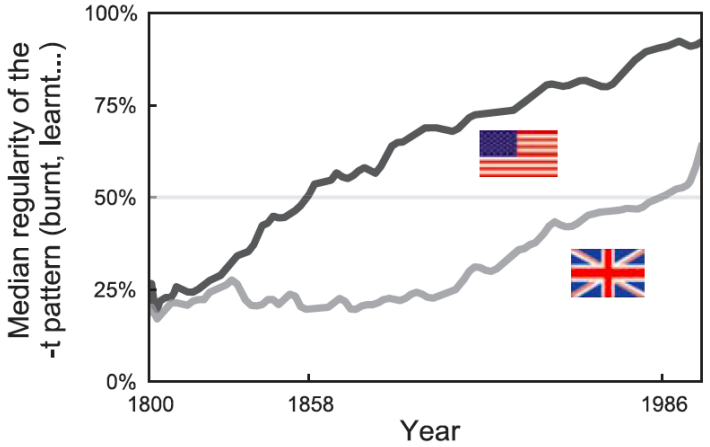
- burn,
- chide,
- smell,
- spell,
- spill,
- thrive.



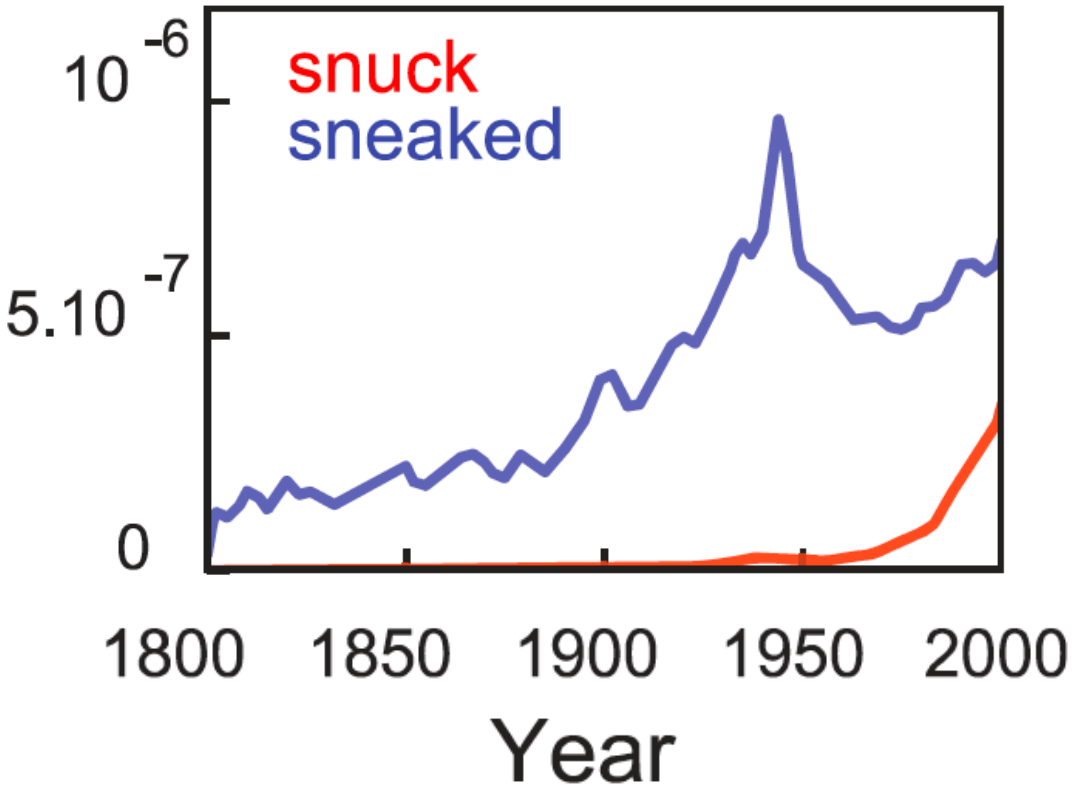
The evolution of grammar: irregular verbs. British vs. American English



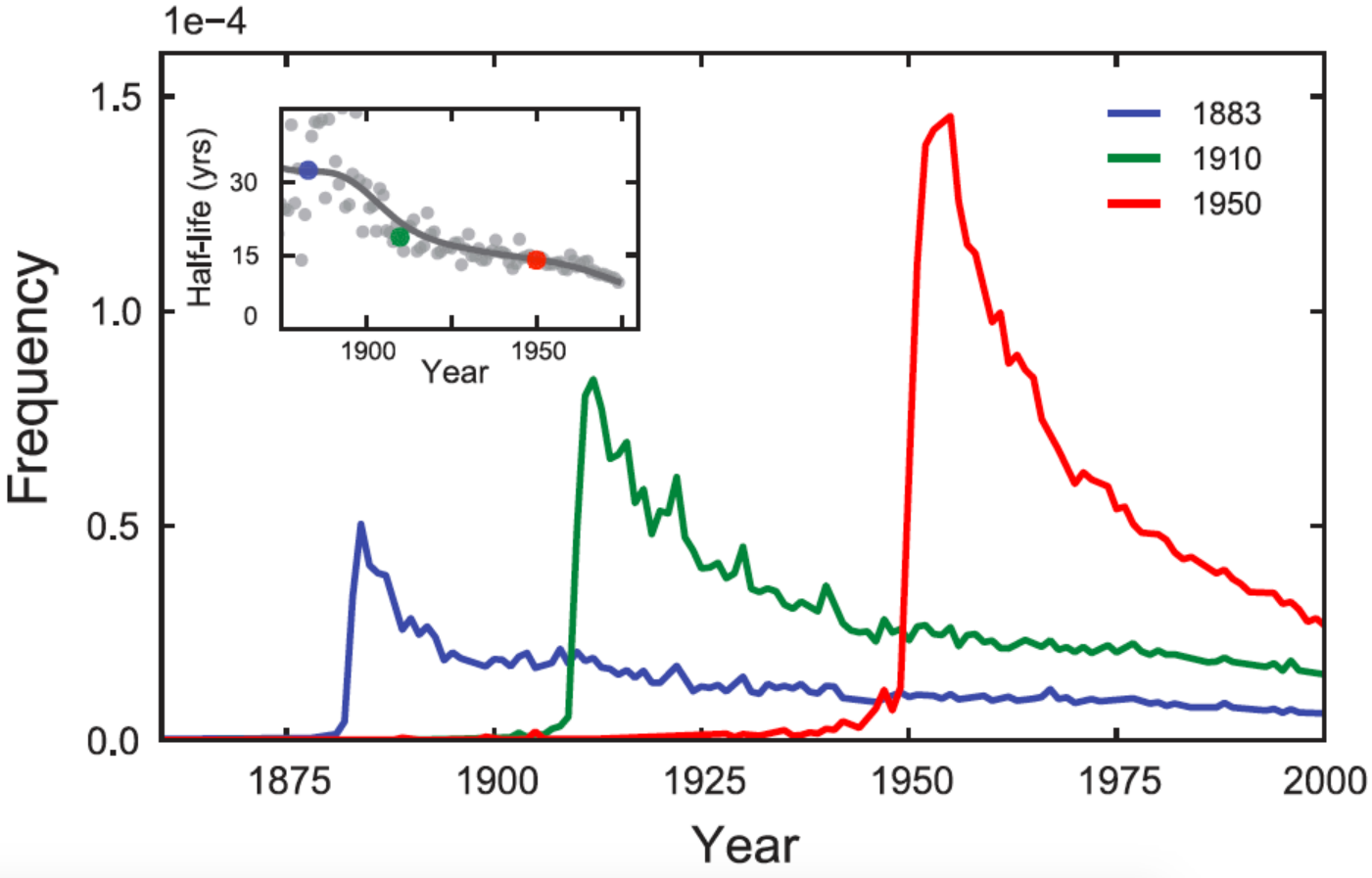




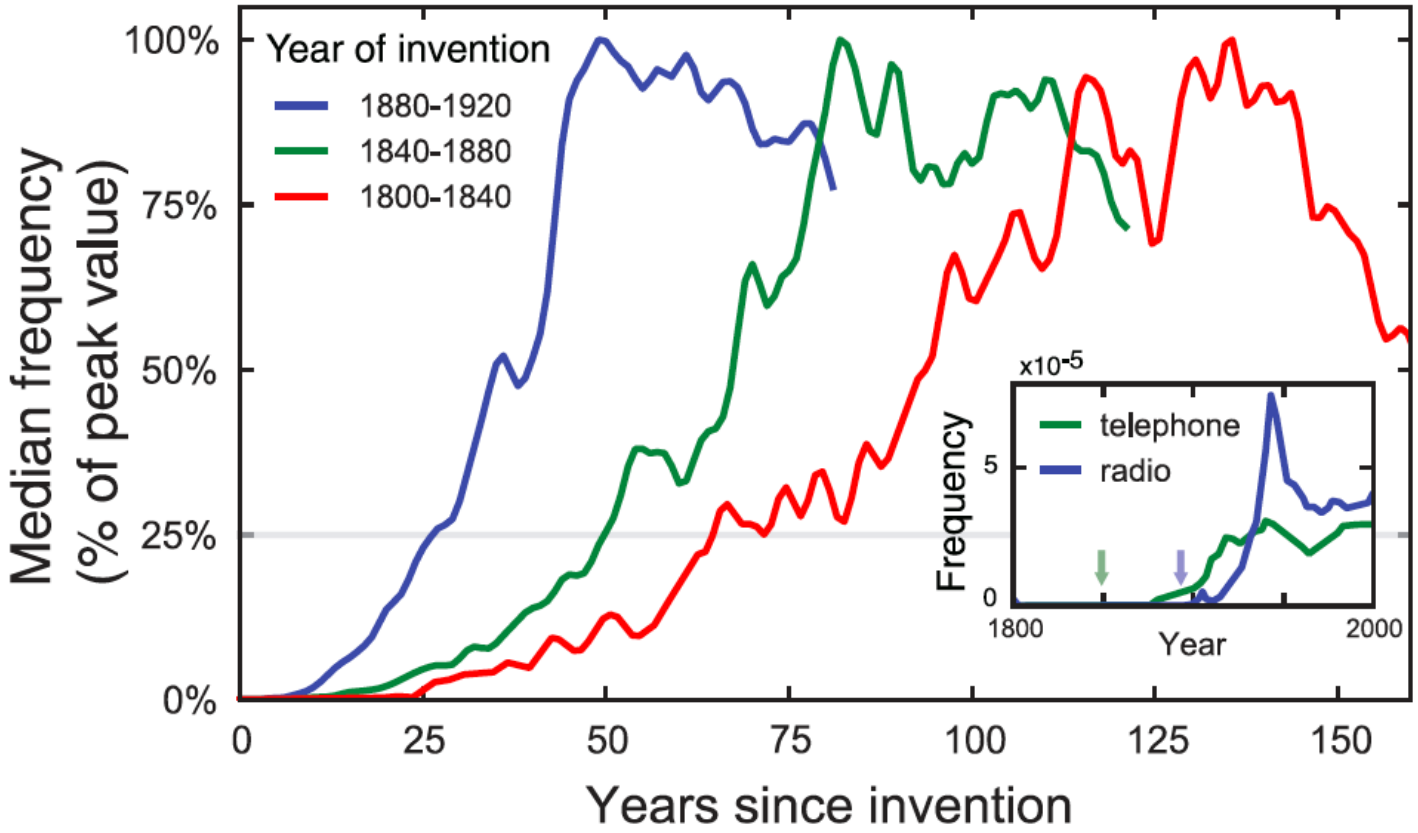
The evolution of grammar: irregular verbs. Irregularization



Forgetting the past



Forgetting the past

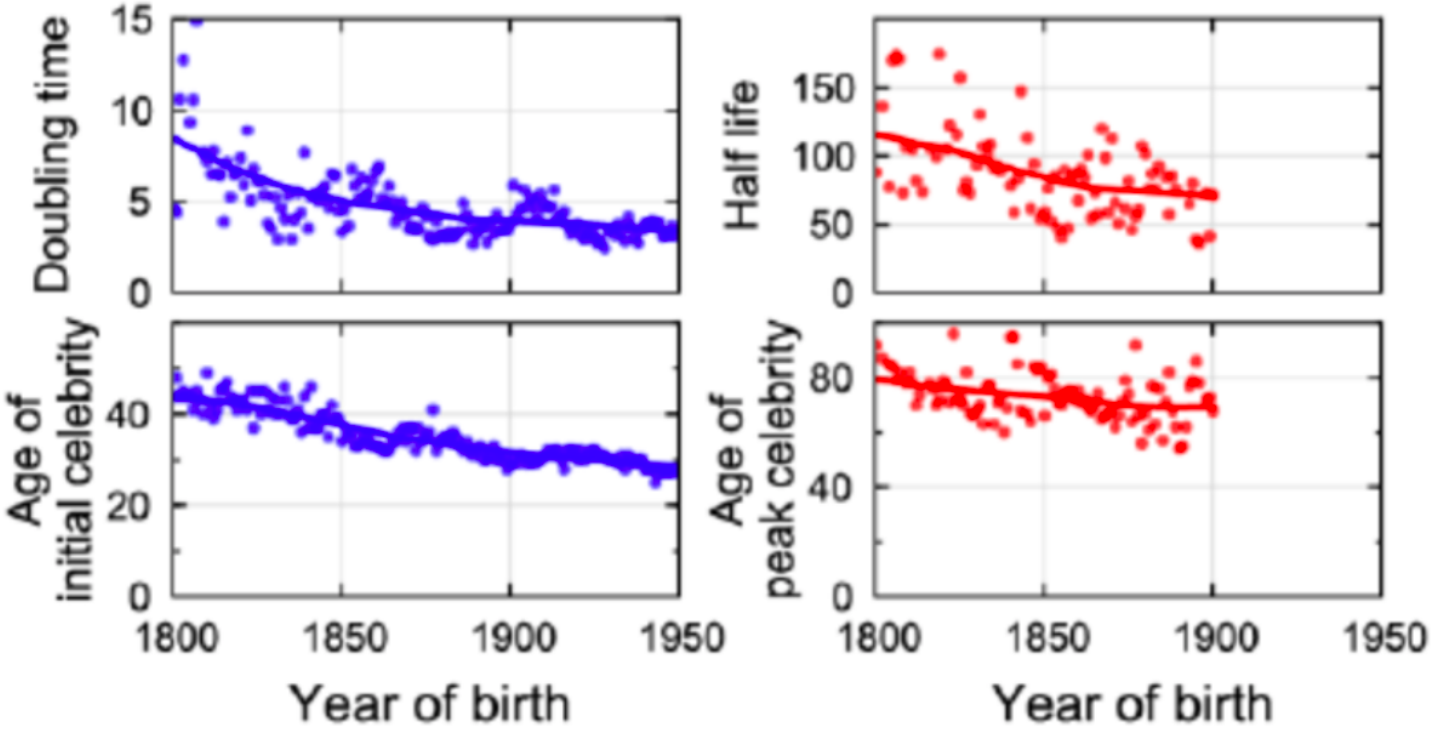


Forgetting people

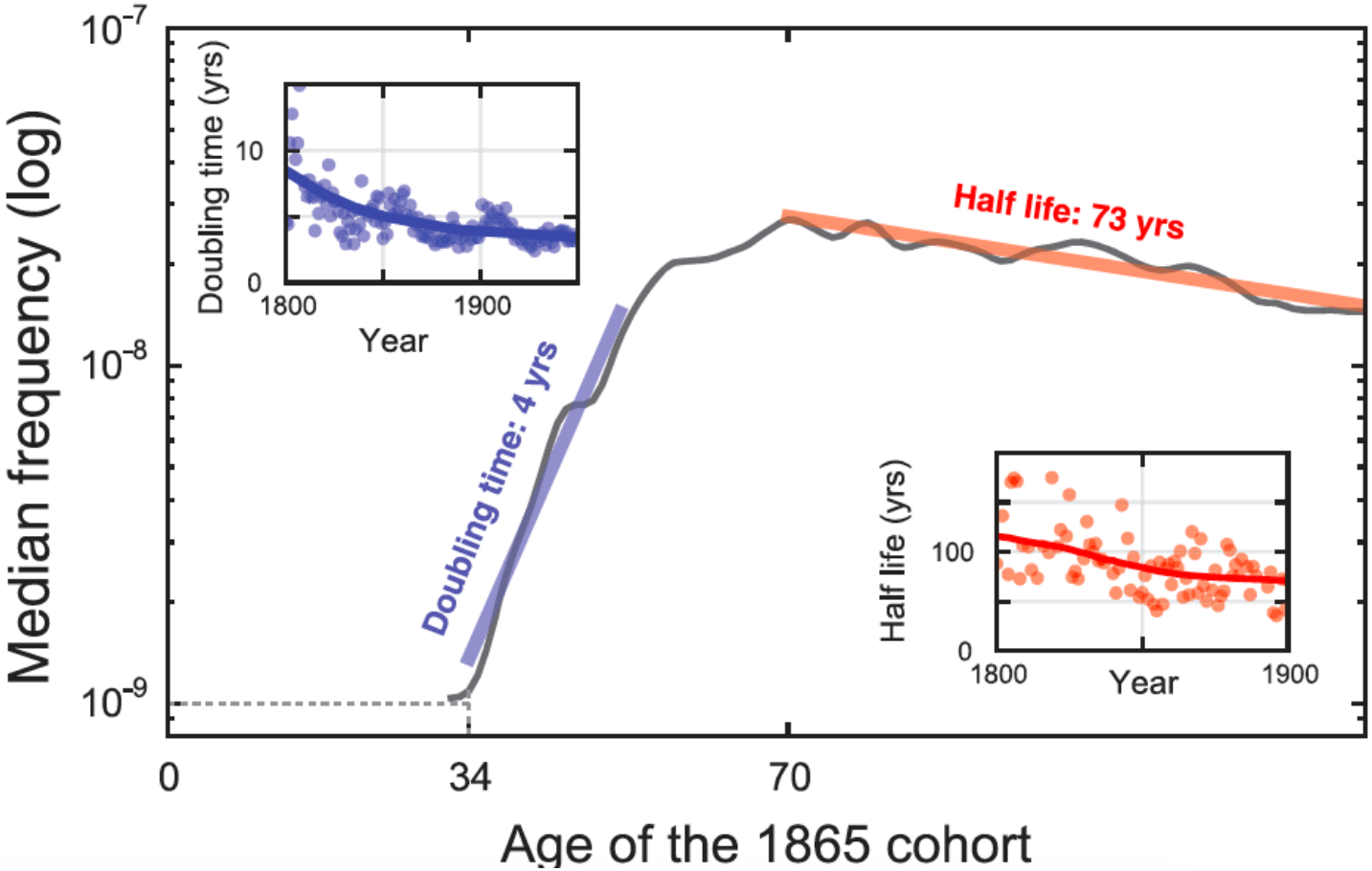
Fame:

- precelebrity period (frequency $< 10^{-9}$) — the age of initial celebrity (43 years in early 19th century and 29 years in mid-20th century),
- rapid rise to prominence — doubling time of initial rise (8.1 years in early 19th century and 3.3 years in mid-20th century),
- a peak — the age of peak celebrity (75 years),
- slow decrease - half-life of the decline (120 years in early 19th century and 71 years in mid-20th century).

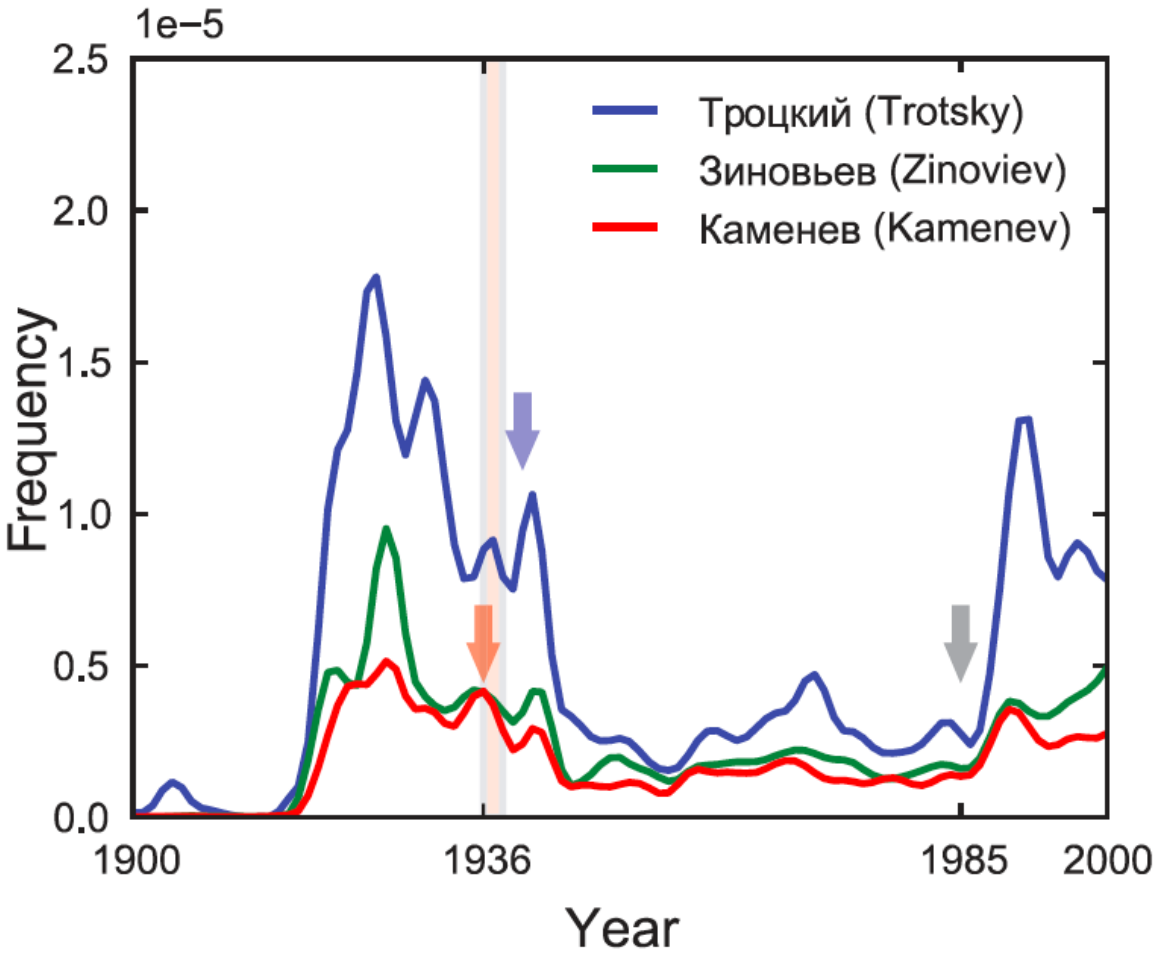
Forgetting people



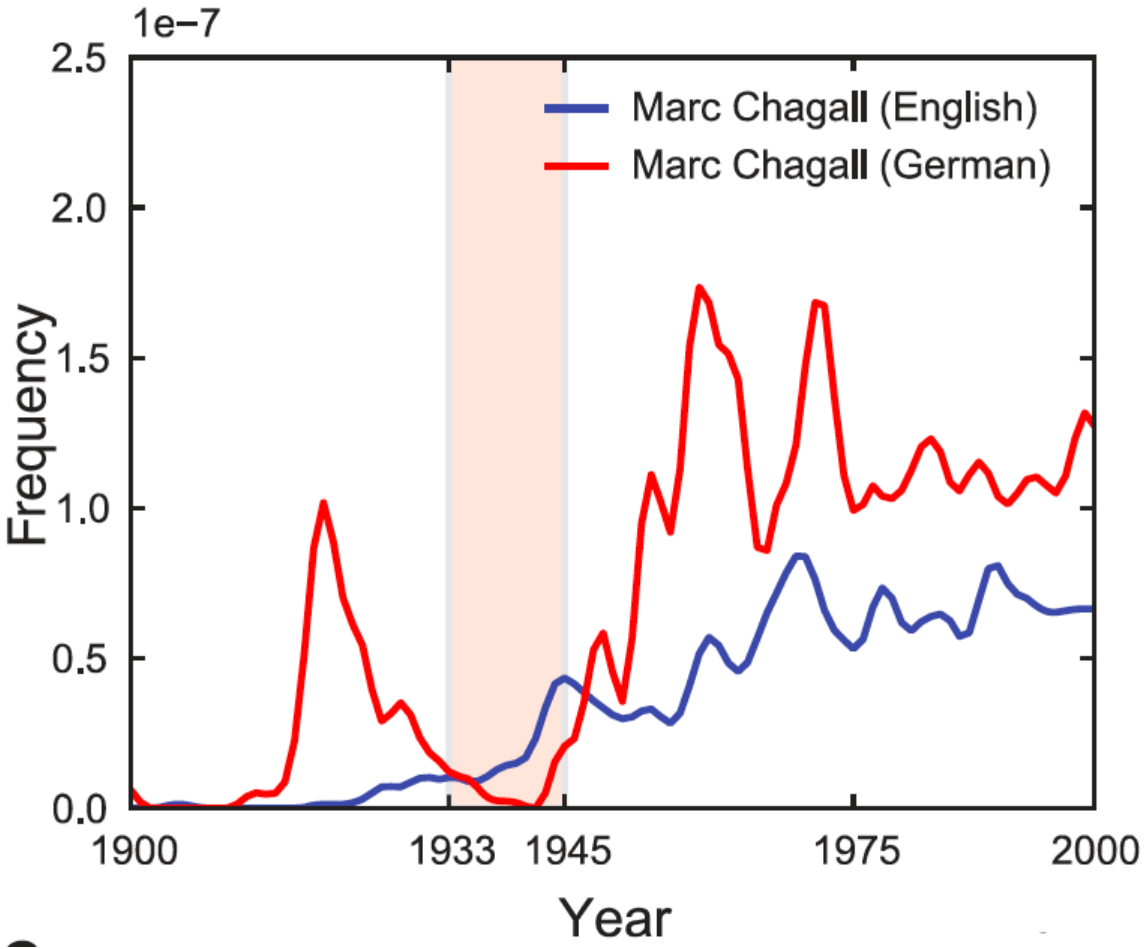
Forgetting people



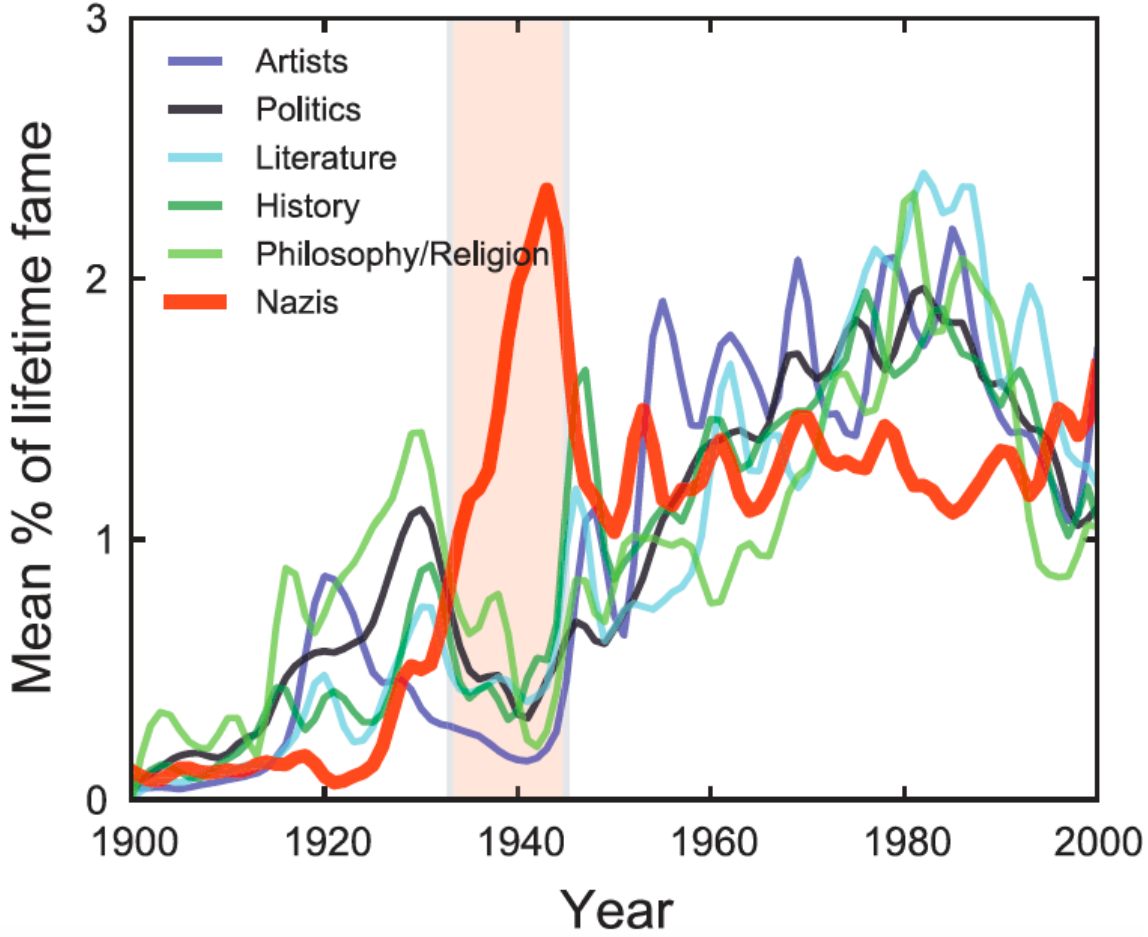
Detecting censorship and suppression



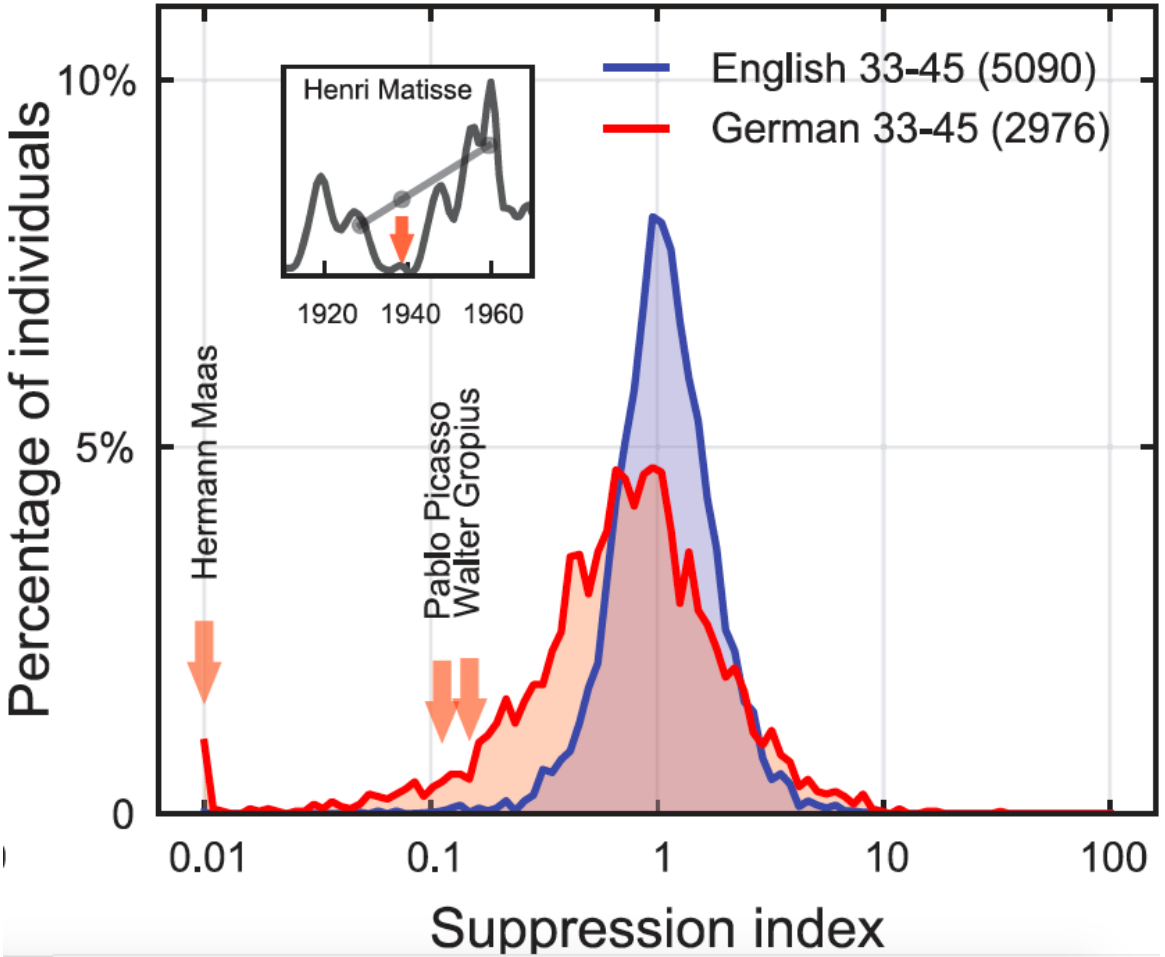
Detecting censorship and suppression



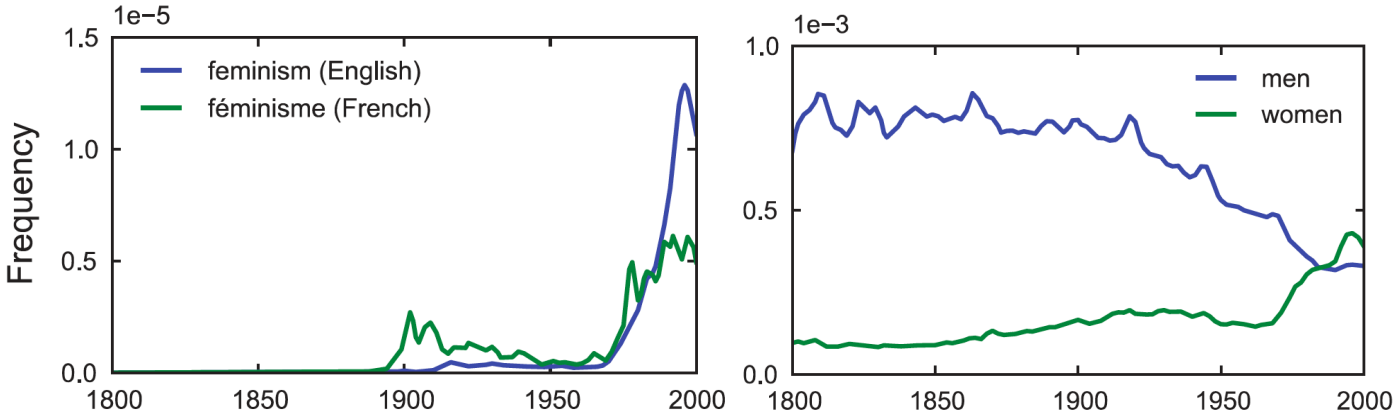
Detecting censorship and suppression



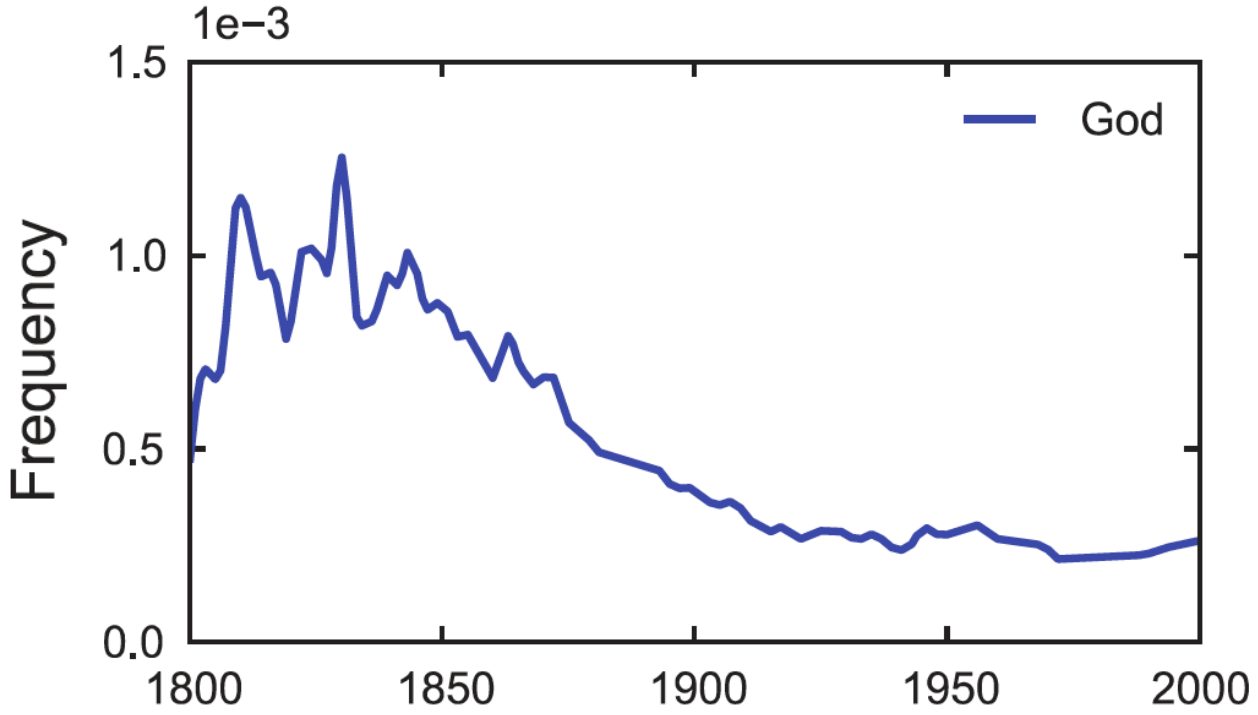
Suppression index



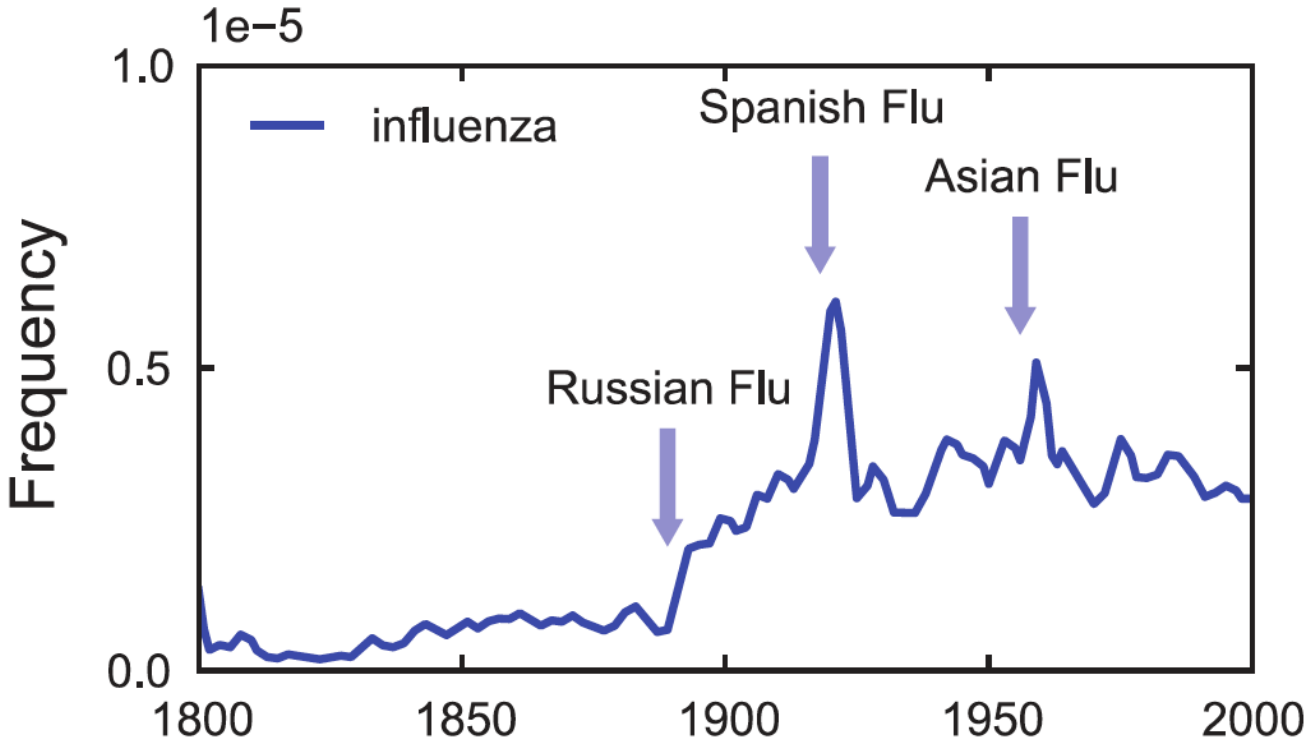
Feminism


1e−5
1.5
1.0
0.5
0.0
Frequency
feminism (English)
féminisme (French)
1800
1850
1900
1950
2000
1e−3
1.0
0.5
0.0
men
women
1800
1850
1900
1950
2000

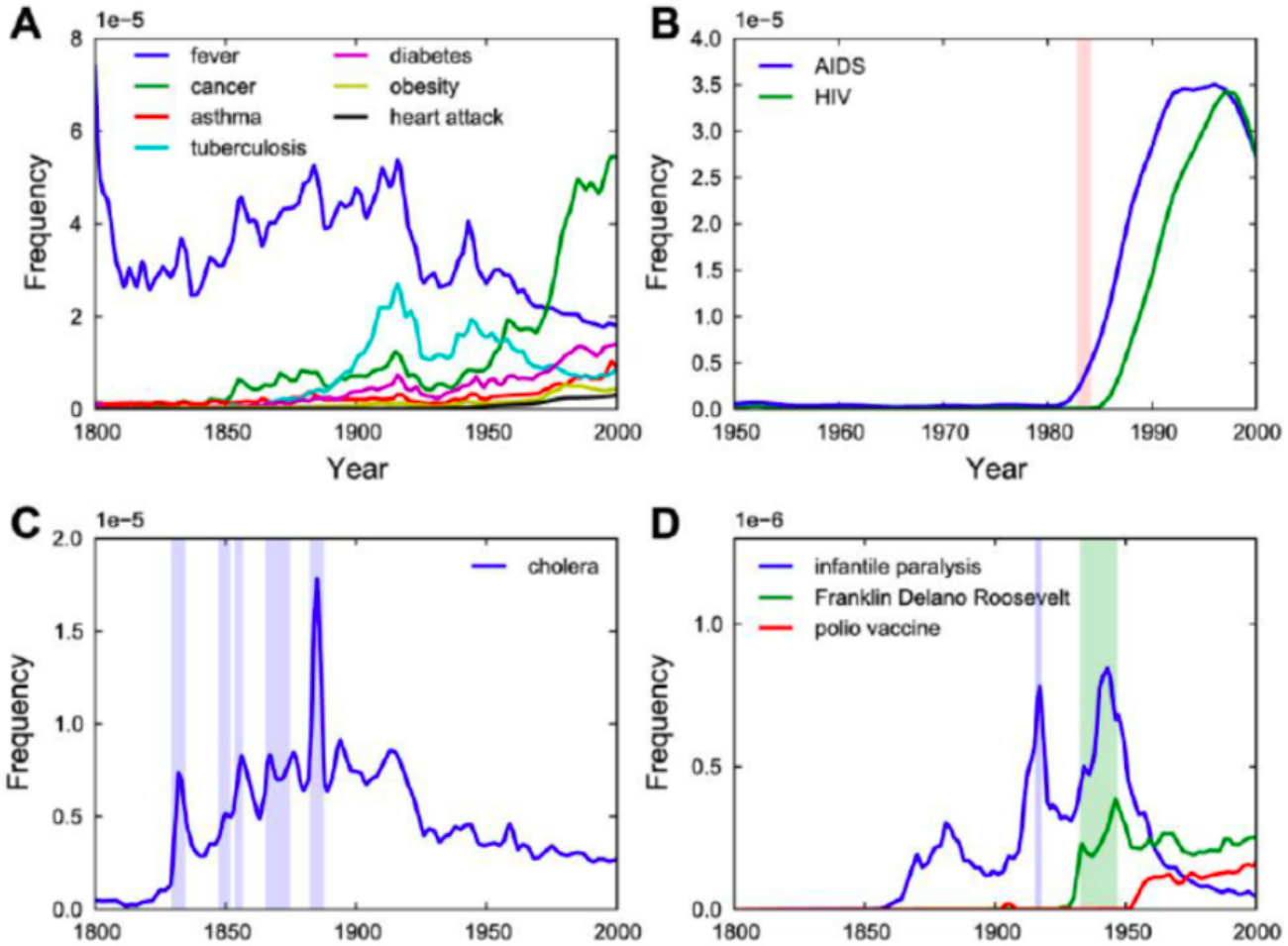
Religion



Epidemiology



Epidemiology



A Multi-Modal Feature Embedding Approach to Diagnose Alzheimer Disease from Spoken Language

Haj Zargarbashi, S. Soroush
s.zargarbashi@ut.ac.ir

Babaali, Bagher
babaali@ut.ac.ir

October 2, 2019

4 Conclusion and Future Works

In this work we used three methods, two of them on the sound signal data (i-vector and x-vector) and one on the sequence of words (N-gram model) for diagnosis of Alzheimer disease and we reaches to accuracy 83.6%. This model can be applied on various languages and even low-resource ones.