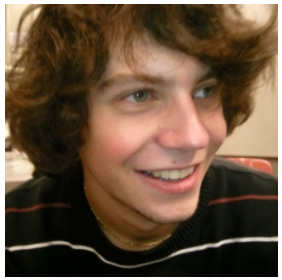
smargn

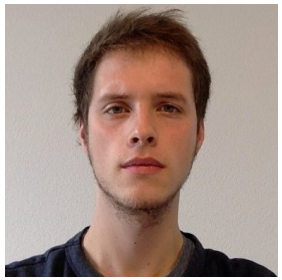
Reversing the n-grams searching process

Z. Gucevska



M. Monney

F. Junker



J. Gaspoz

J. Salathé

A. Manasovska

F. Jolidon

V. Rutz

S. Bovet

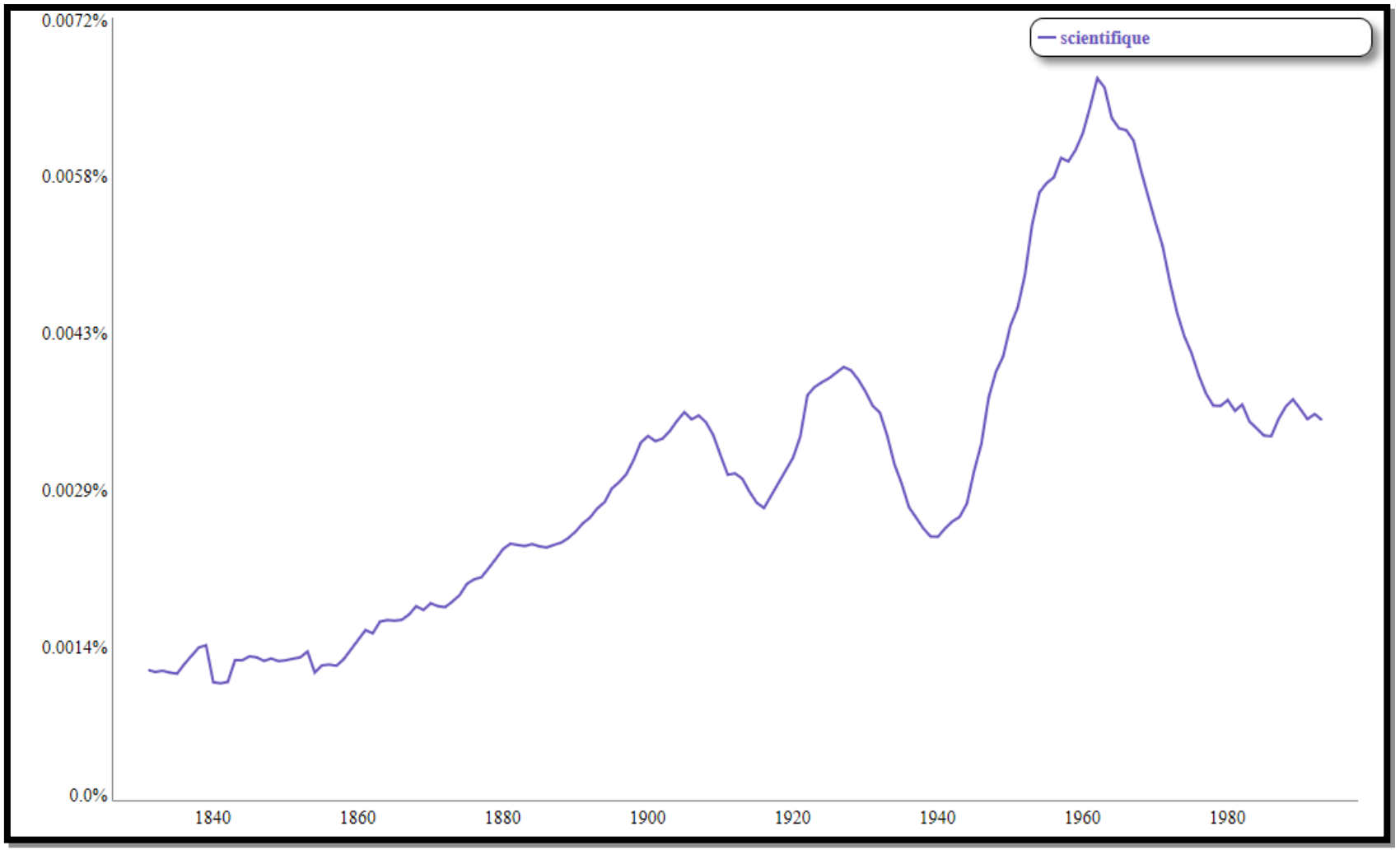
Team

- Le Temps newspaper
  - La Gazette de Lausanne
  - Le Journal de Genève

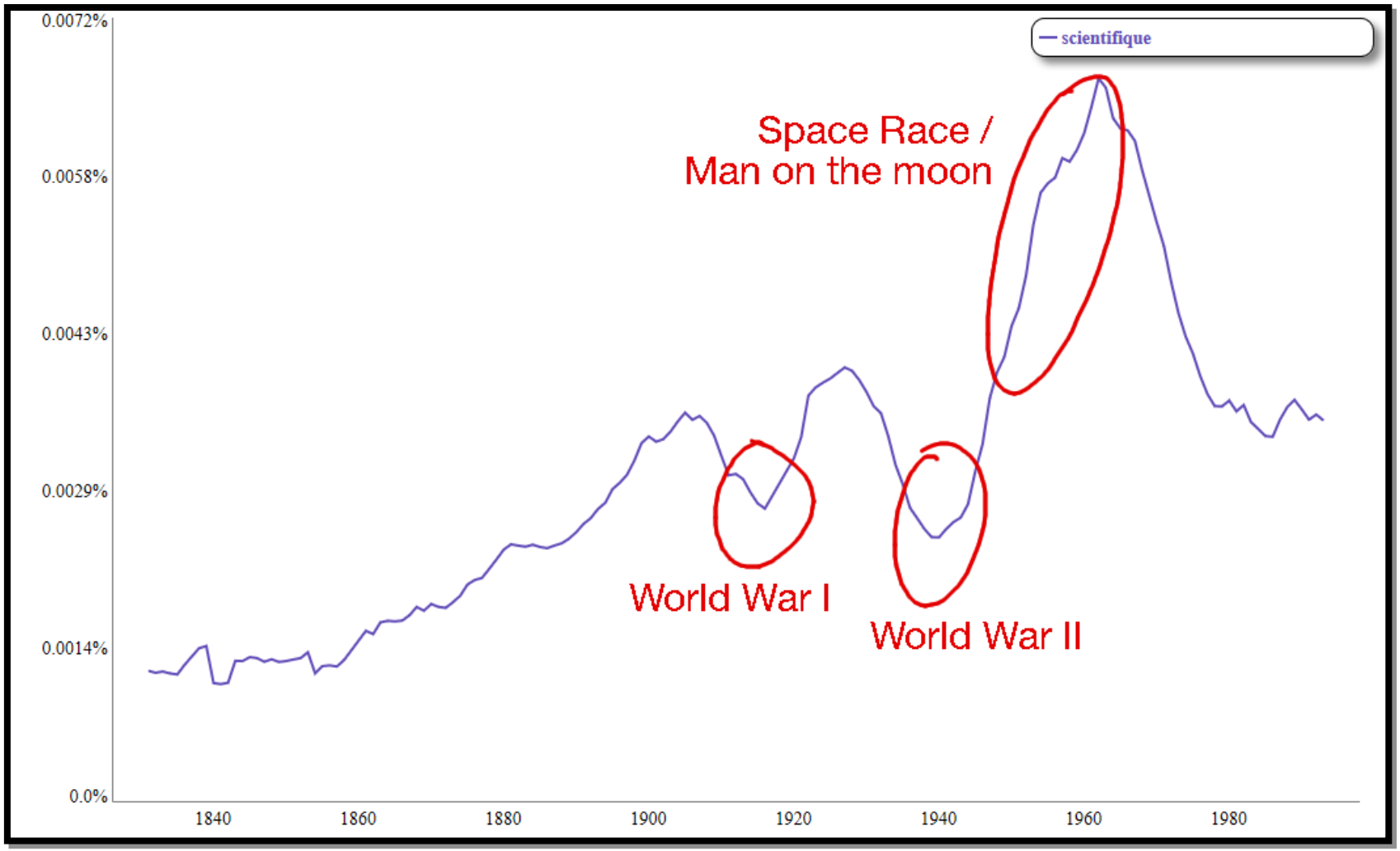
- 159 years of articles (~16 GB)

- Provided by *DHLab* in association with *Le Temps* and the *Bibliothèque nationale suisse*

Corpus



Understand a profile



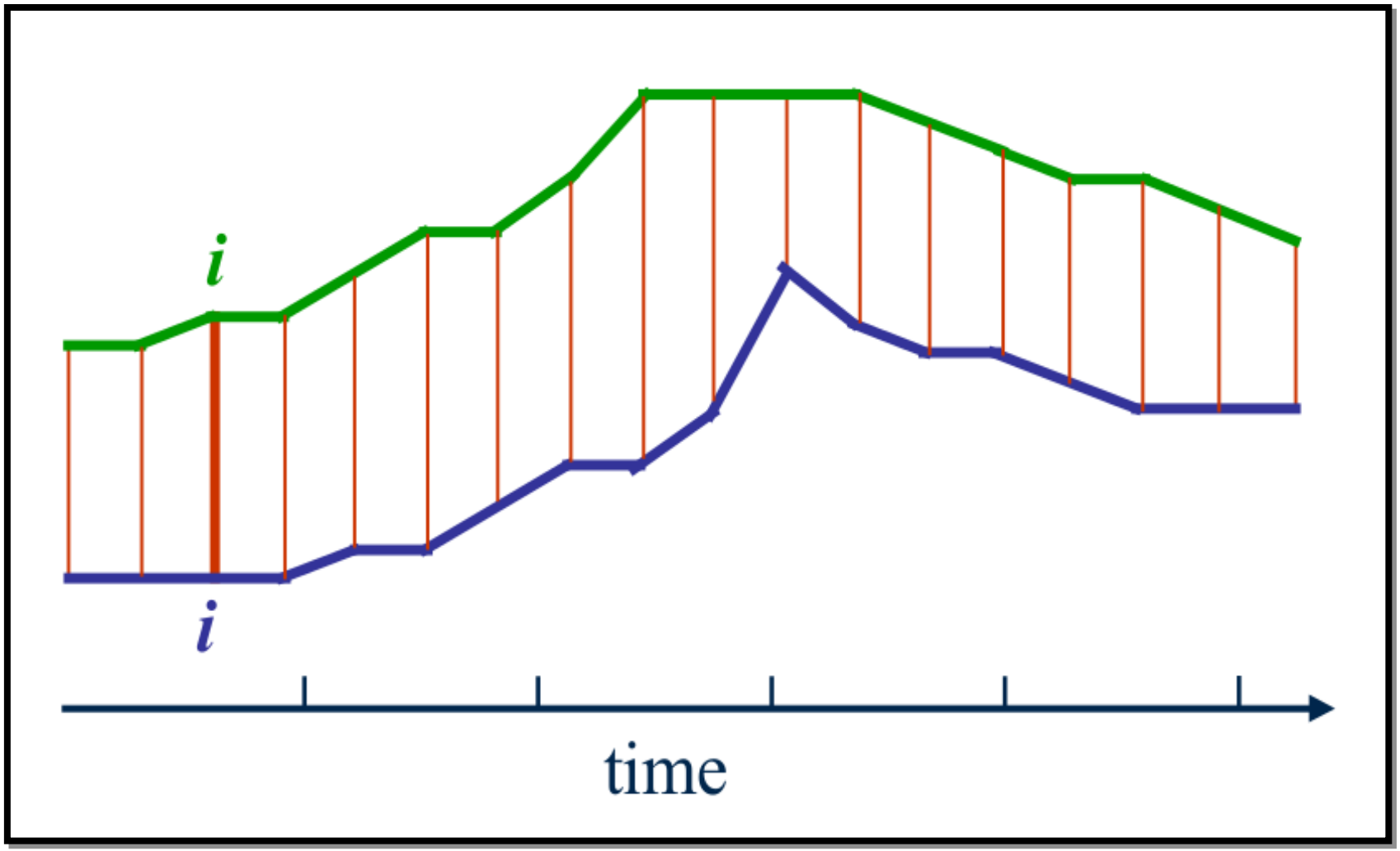
Understand a profile

- Studies based on n-grams have become popular
- Most of them relied on researcher's intuition
- We provide a way to reverse the process
  - Look for interesting words
  - Find similar temporal profiles
  - Understand why

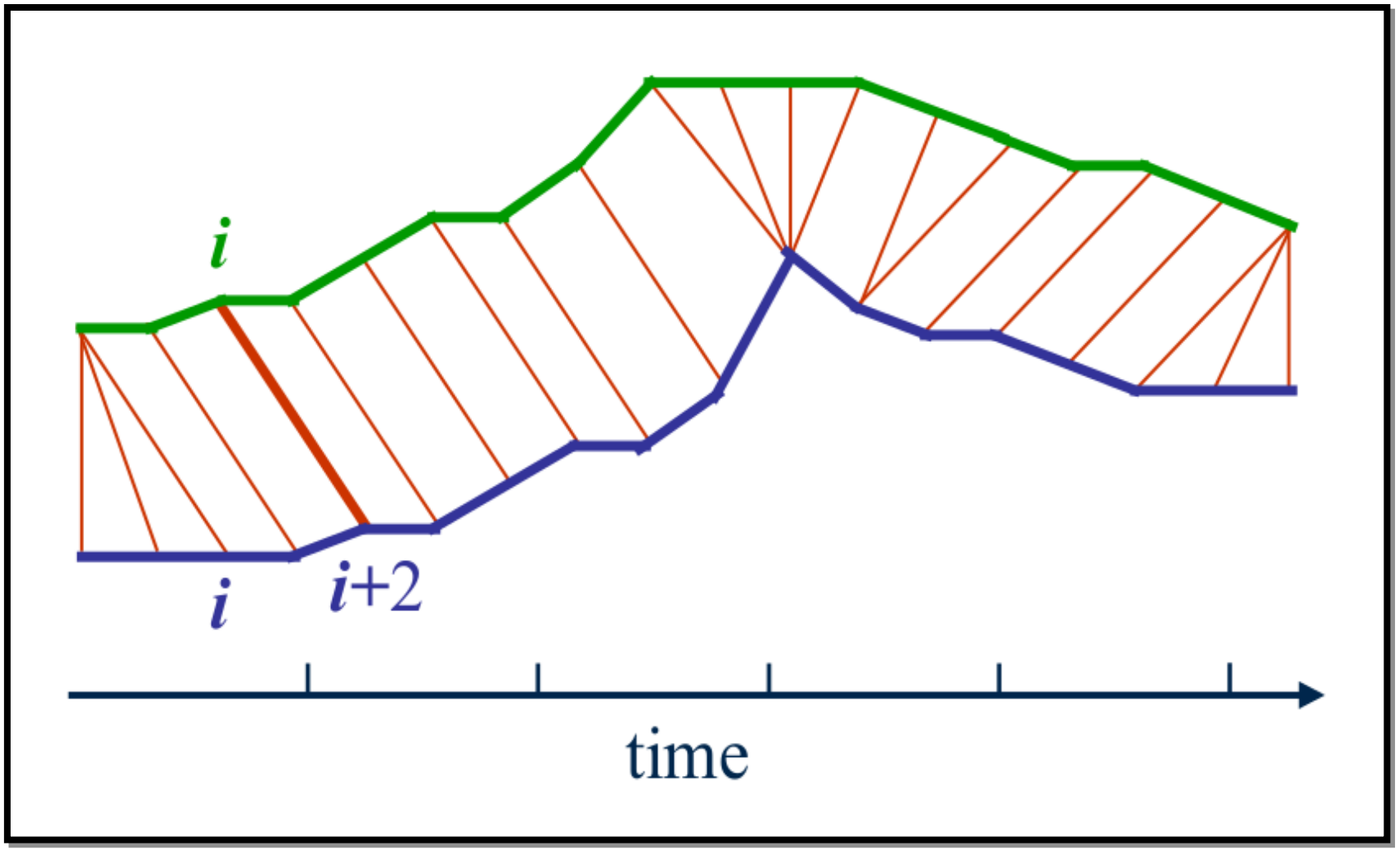
Motivation

- Different comparison techniques implemented
  - Naive comparison
  - Dynamic Time Warping
  - Peak detection
  - (many other unsuccessful attempts)

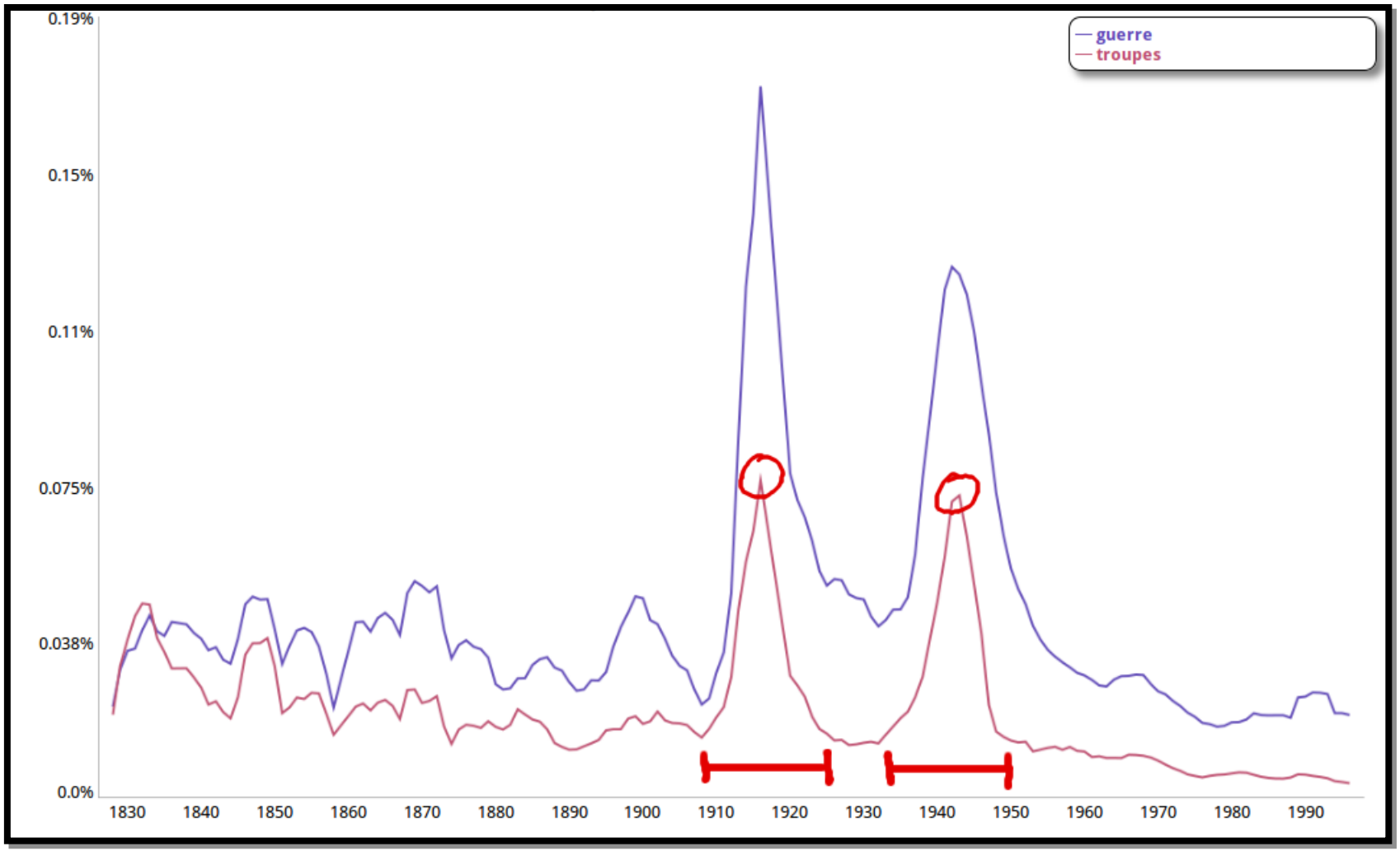
How smargn works



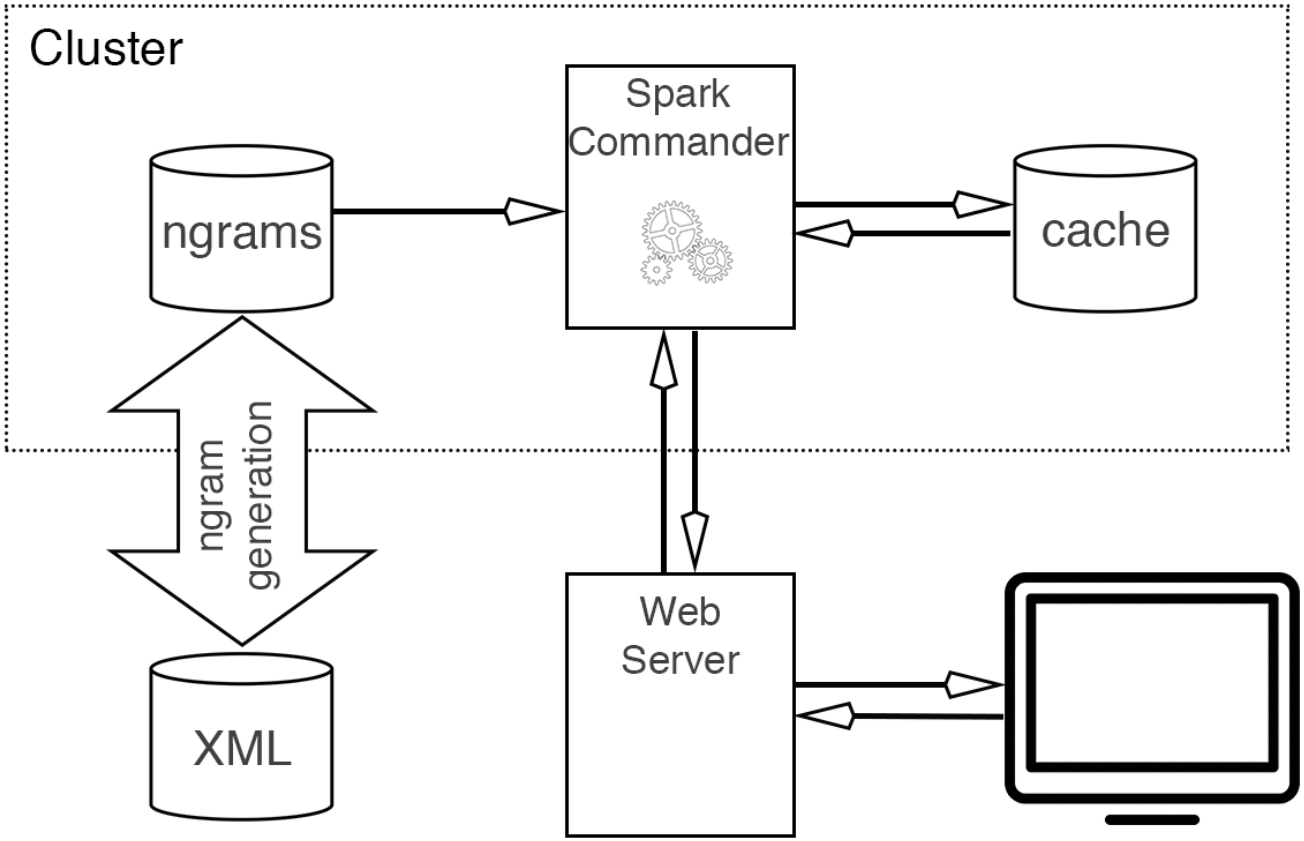
Naive comparison



Dynamic Time Warping



Peak detection



Software architecture

Smargn seach for similar words

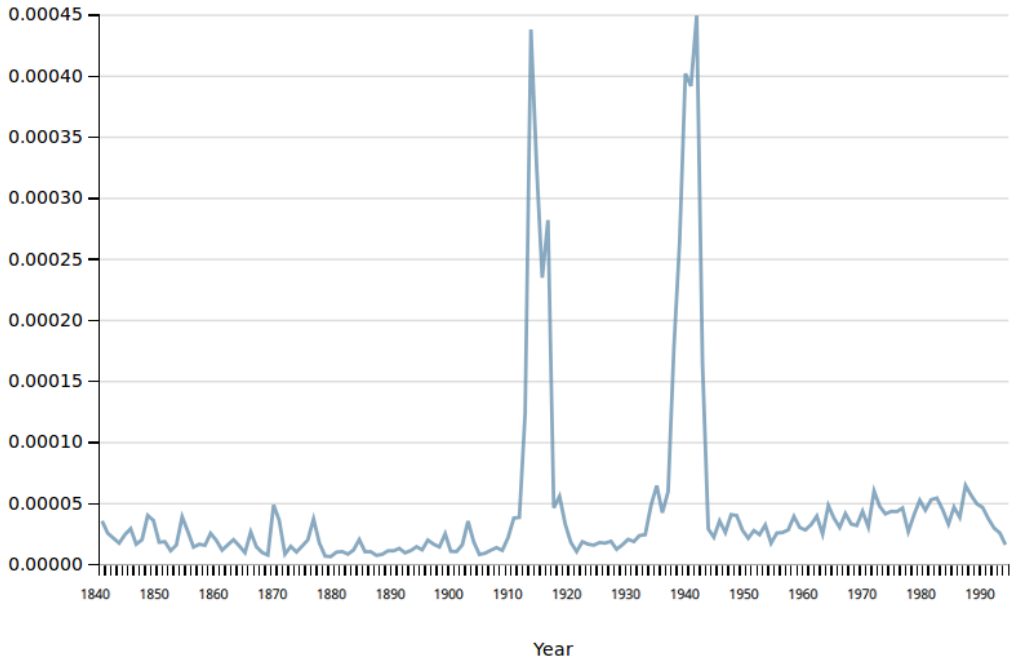
combats

word2

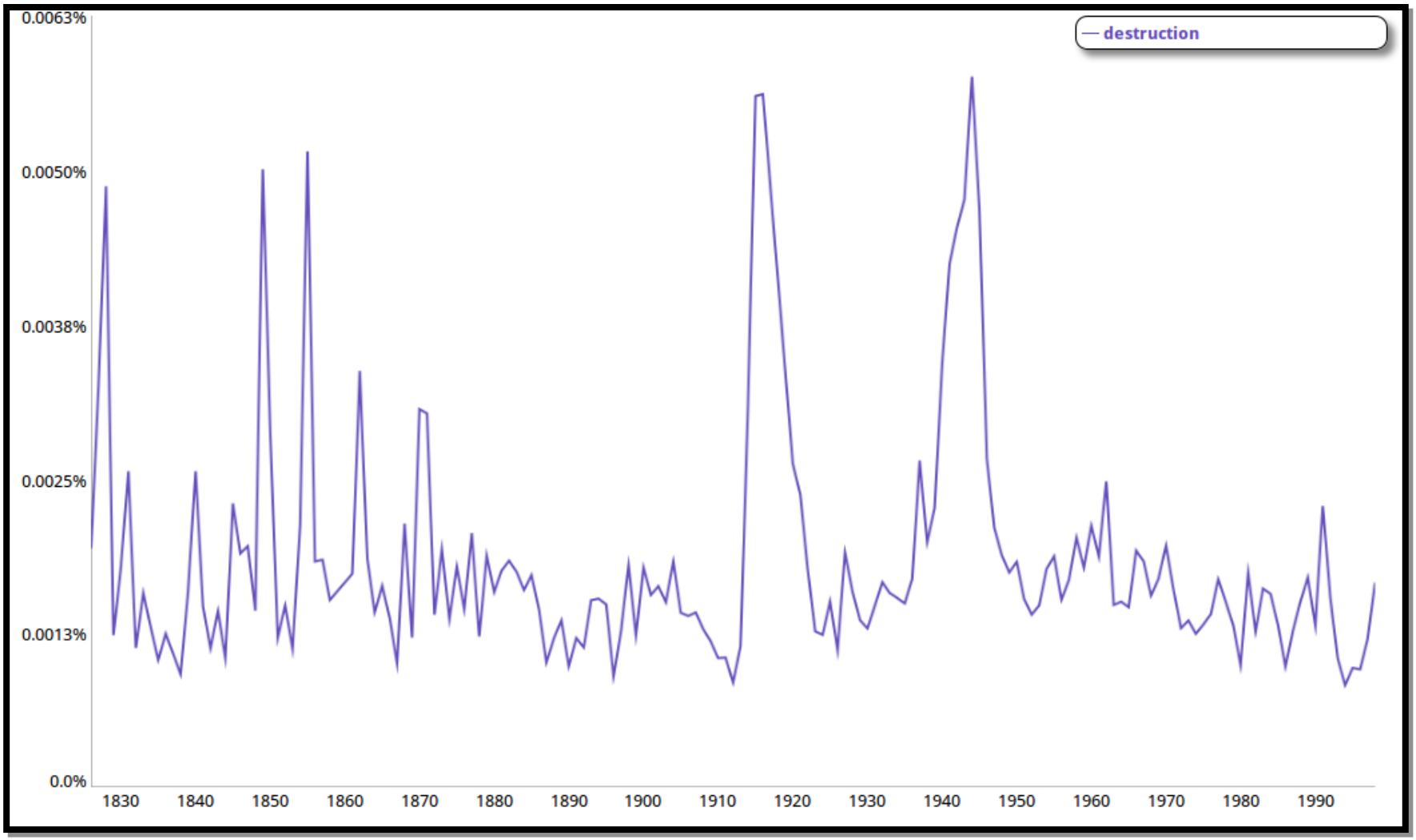
word3

word4

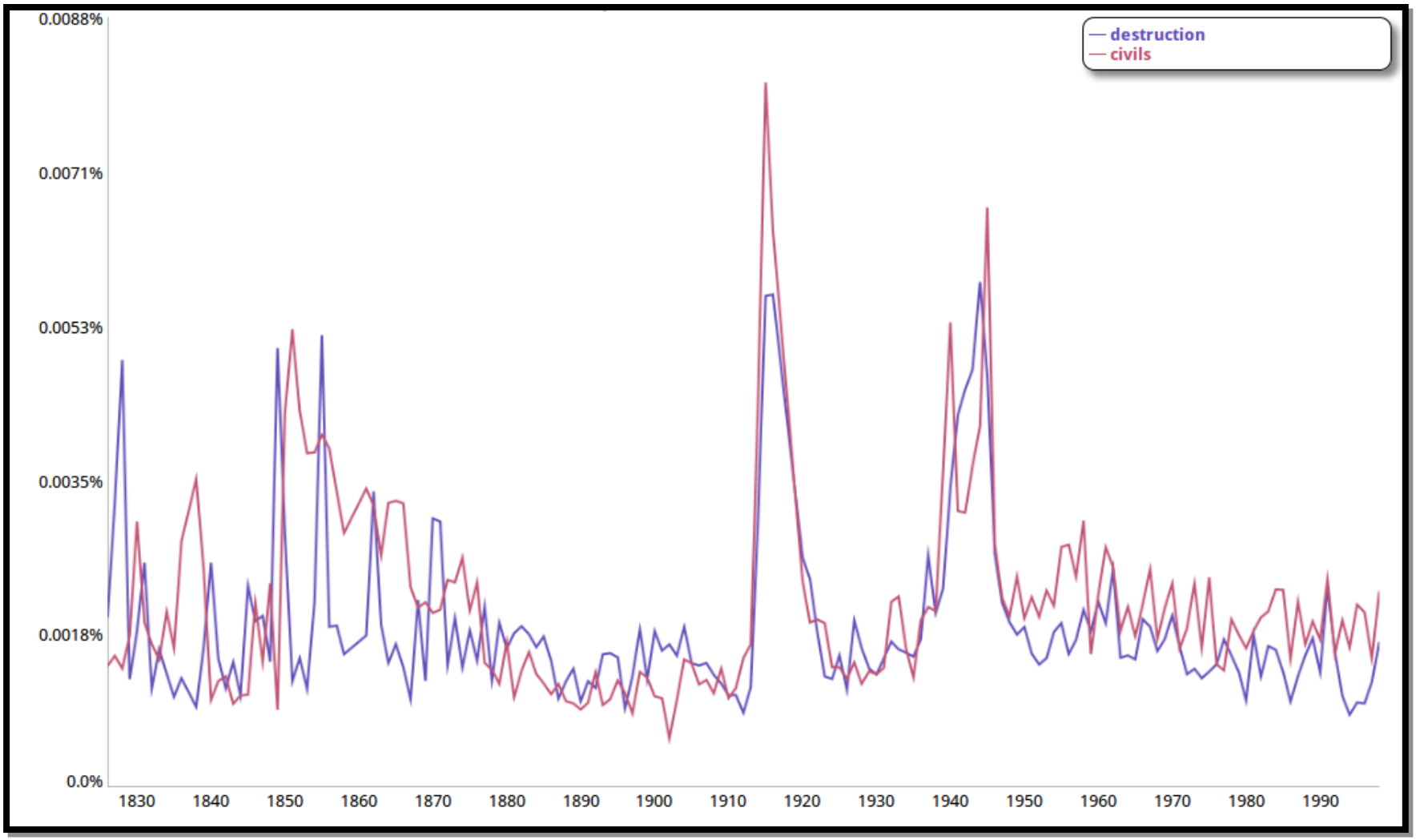
Smargn 'em all!



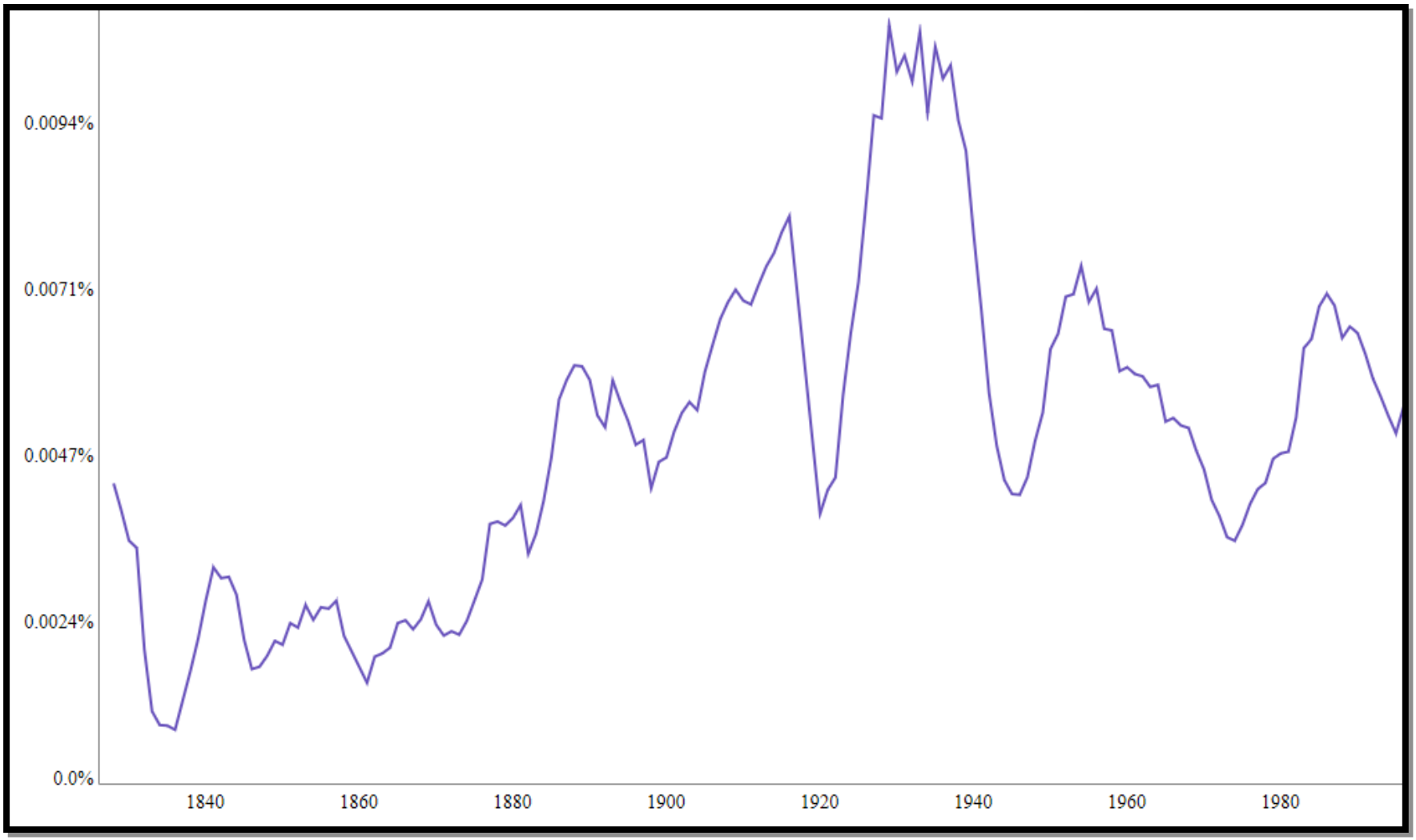
Web interface



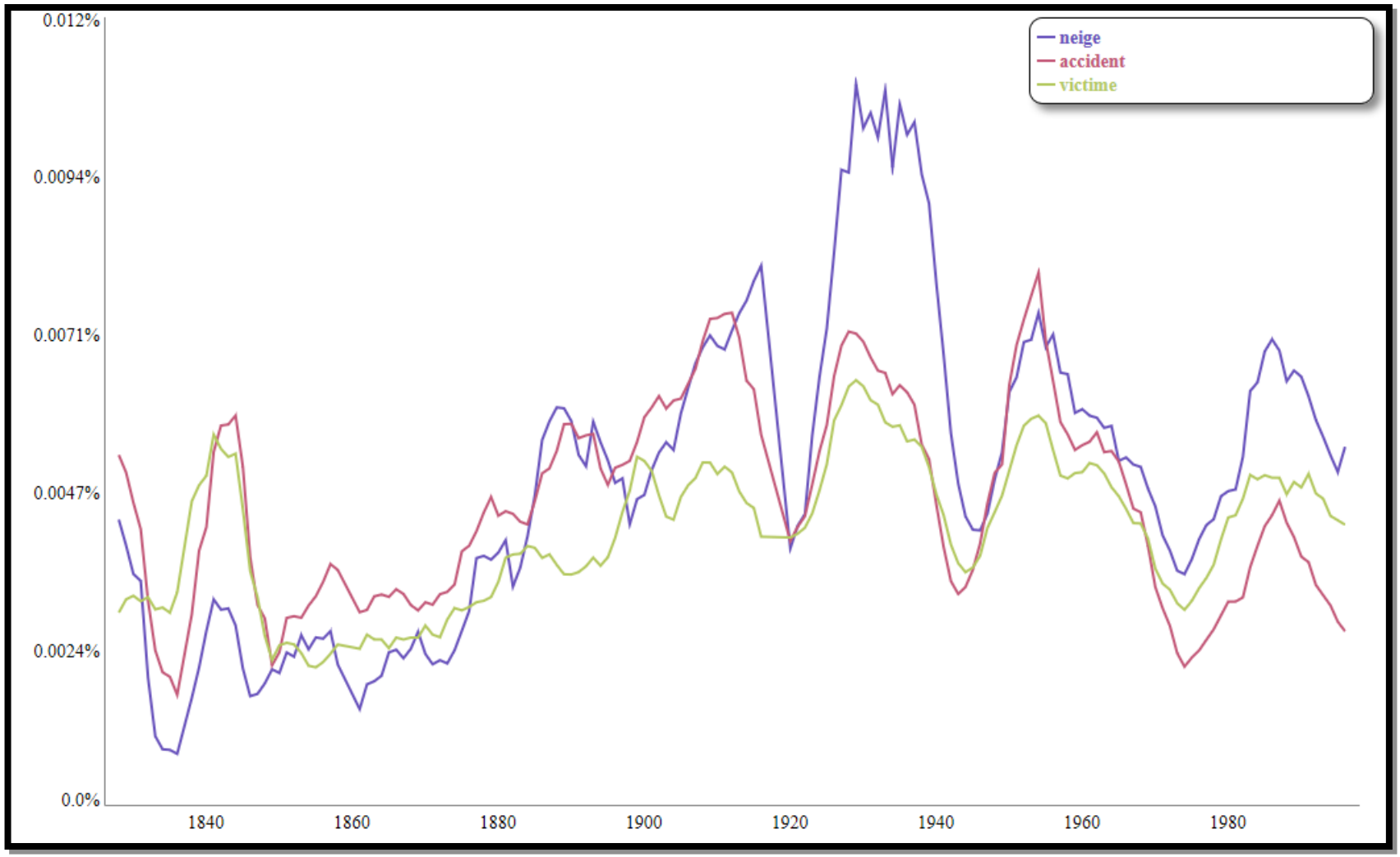
destruction



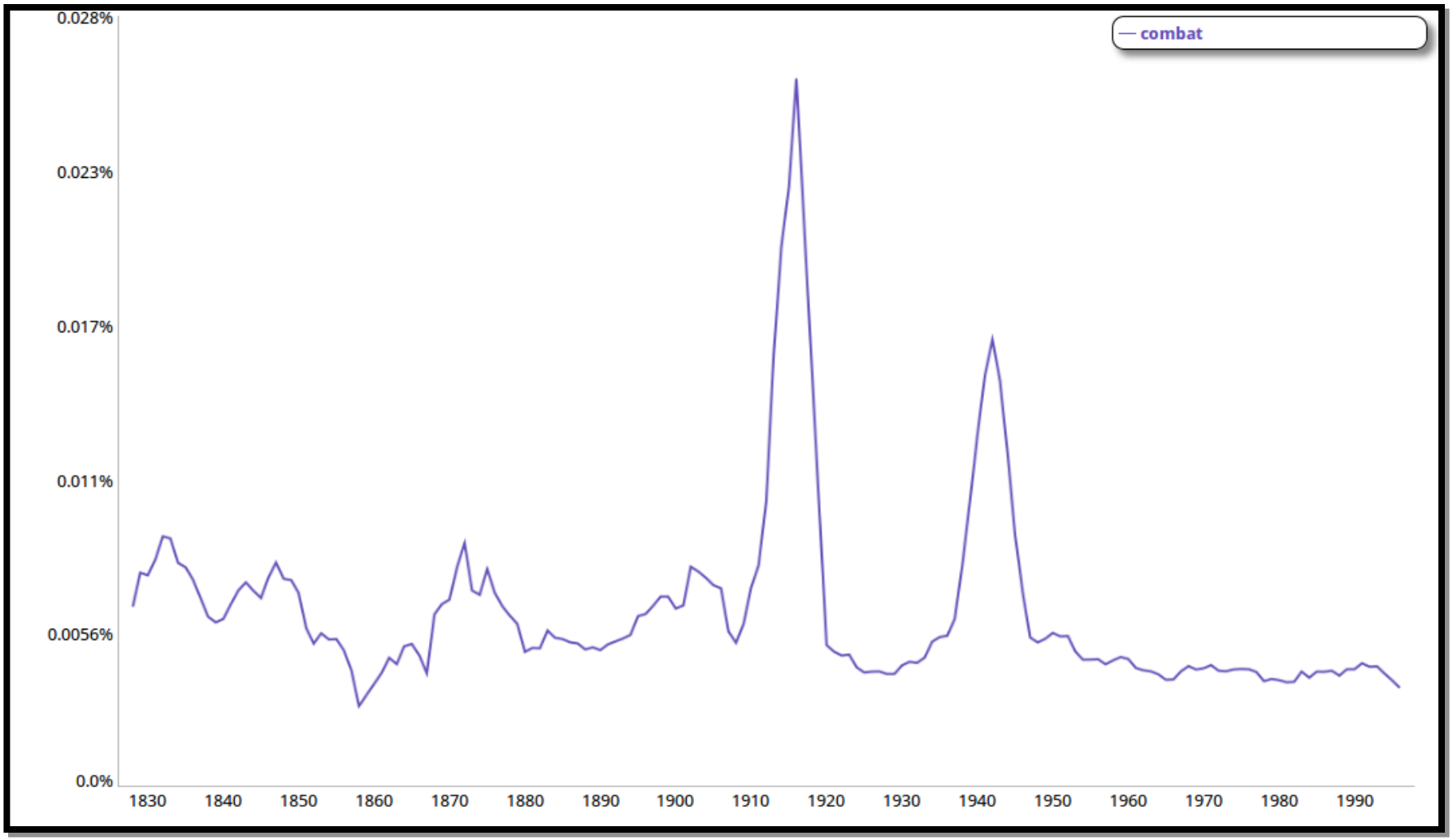
destruction & civils


0.0094%
0.0071%
0.0047%
0.0024%
0.0%
1840
1860
1880
1900
1920
1940
1960
1980

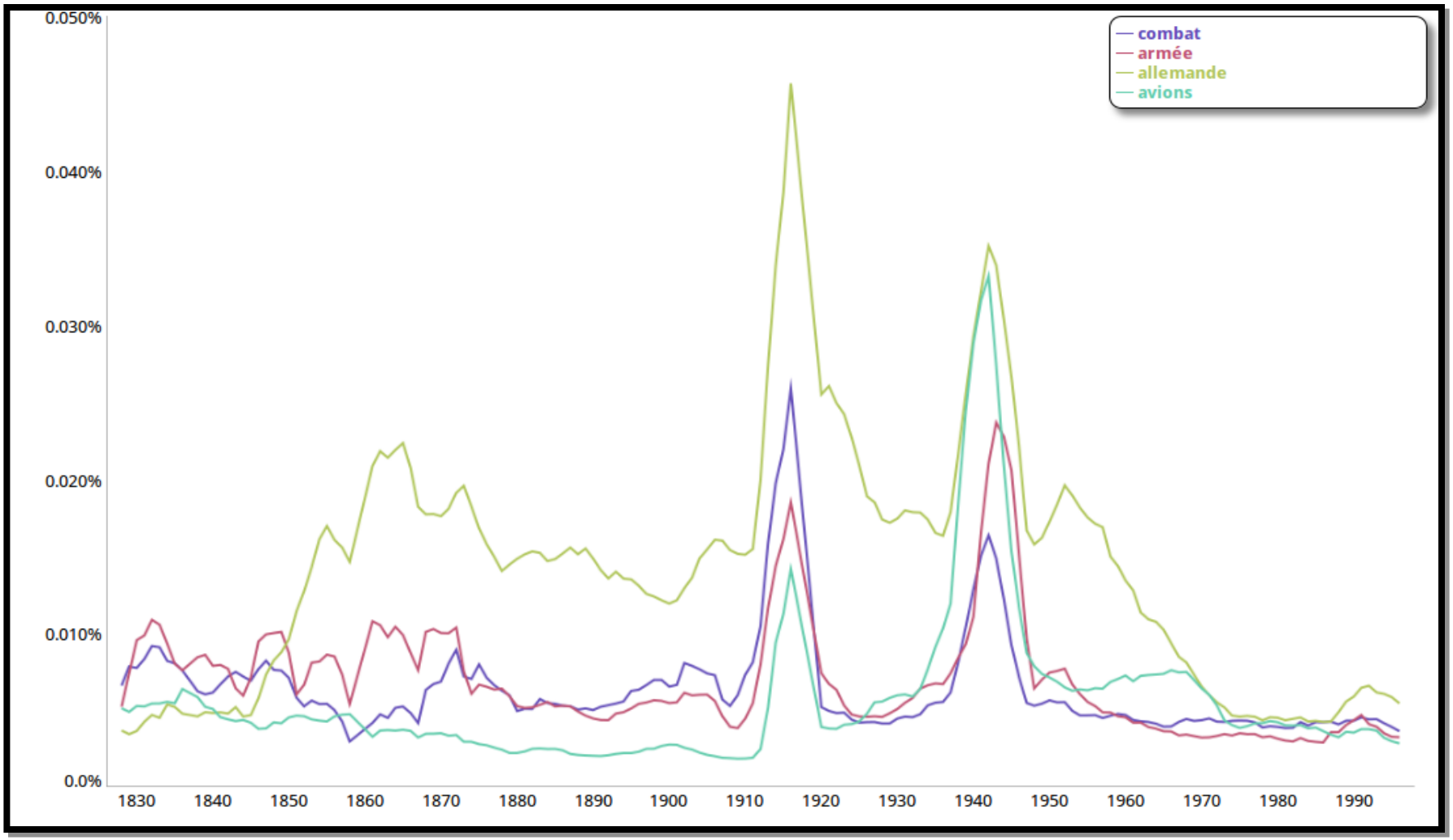
neige



neige, accident & victime 16


combat
0.028%
0.023%
0.017%
0.011%
0.0056%
0.0%
1830
1840
1850
1860
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990

combat

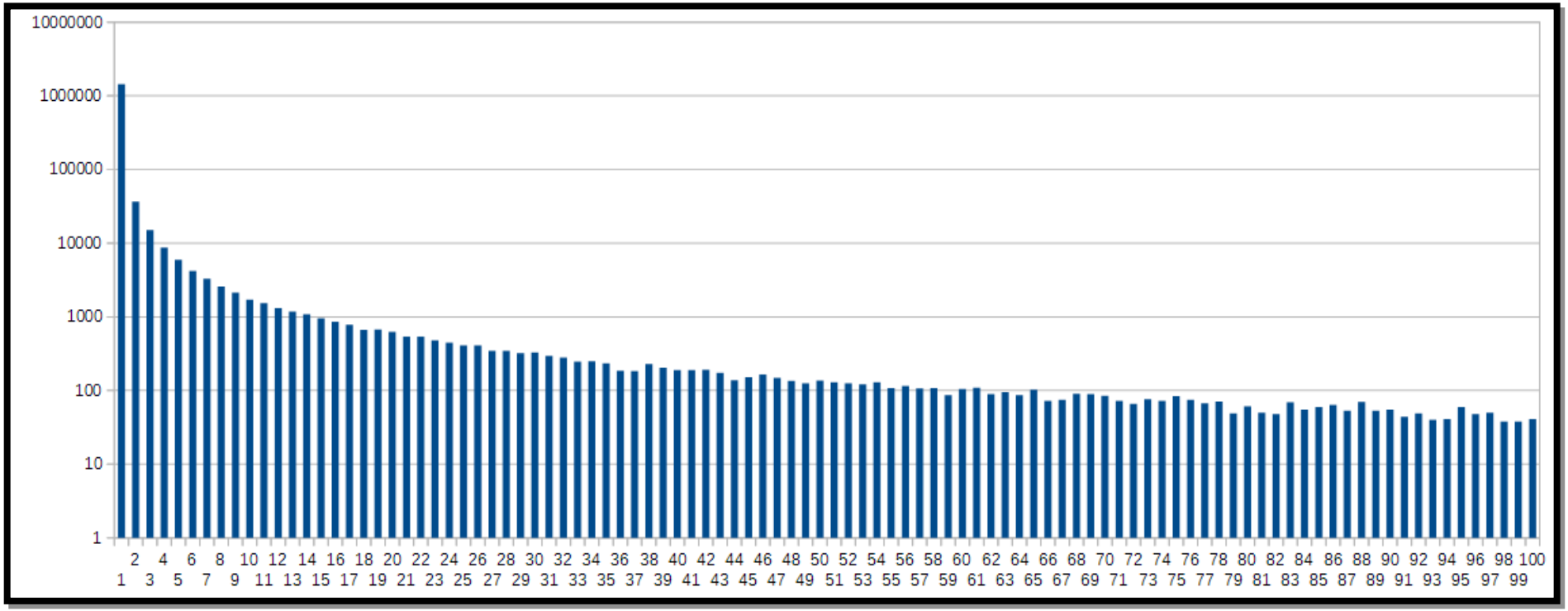


combat, armée, allemande, avions

- Can greatly help linguistic and historical studies
- Gives hints on where to look at
- Makes sure no words are forgotten

Conclusion

Thank you!



Data histogram