DH4CL'ers, Exercise 1

John Nerbonne
Rijksuniversiteit Groningen &
Albert-Ludwigs-Universität, Freiburg

Tübingen
WS 2023

Background

- Culturonomics, TedTALK
- Google collected millions of books, originally used for language modeling
- Michel & Lieberman headed exploratory team
  - Do frequencies of words (and n-grams) reflect culture at their time (of printing)?
  - https://books.google.com/ngrams/
- Suggestive of how large text might be interesting to DH
- Easy to use, entertaining

… the inevitable criticisms

- Choice of texts is uneven
  - Statisticians: CONVENIENCE SAMPLE
- Quality of scanning, preparation often lacking
- Note importance of smoothing to obtain clearer views of tendencies
- No translations available, but you can search in other languages (see exercise).

See exercise

- Exercise on moodle
- Note chance (and need) to disambiguate
  - E.g. *black* as adjective, common noun
  - Often difficult to apply
- Note need to include more than one word (or n-gram) to reflect some trends
  - *Slavery, slave, slaveholder*
  - *Colonial, colonialism, colony*
- Find an interesting comparison, write it up (in max. 200wd.) and turn it in via email within one week (Oct. 31)

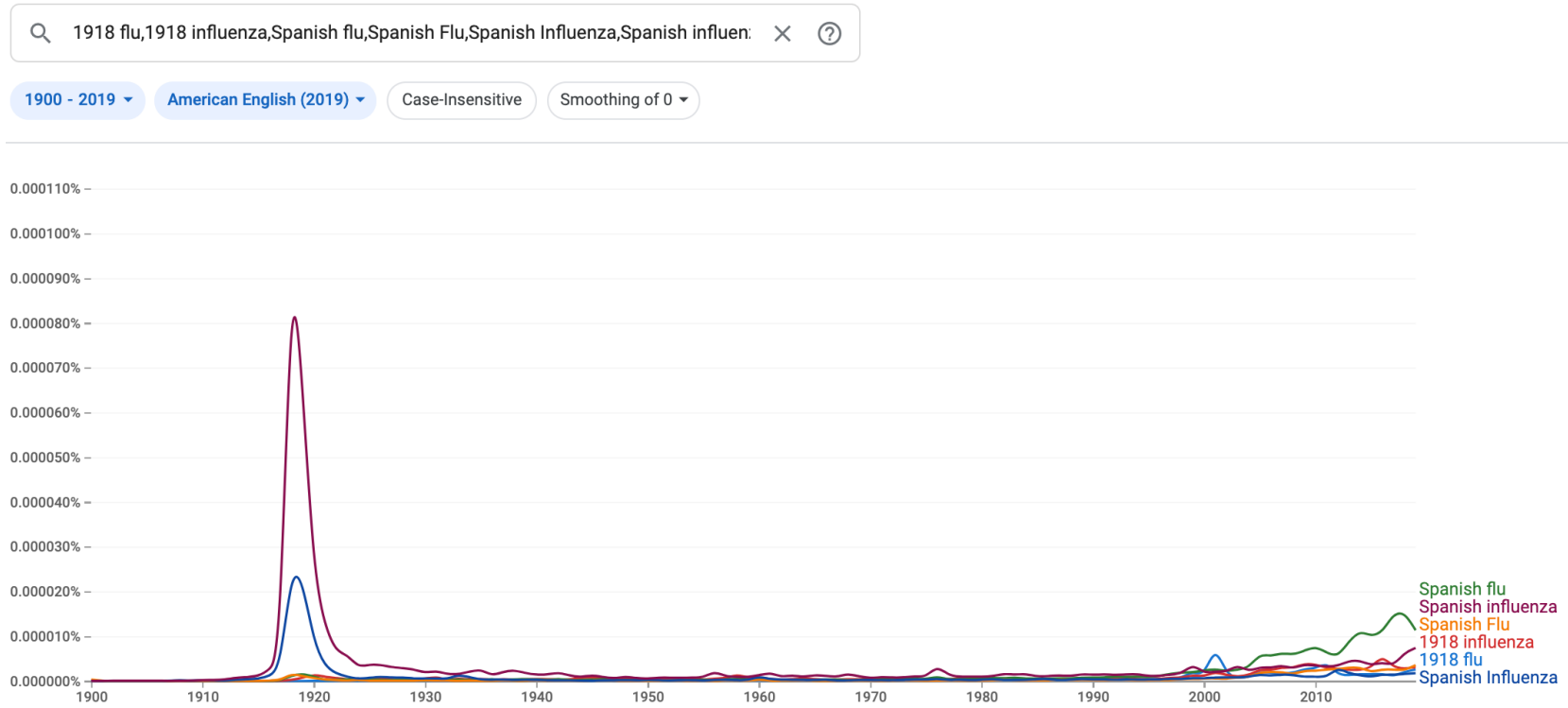
Extra (optional)

- Offer to debate the thesis:
  - **Large-scale text analysis reflects the cultural attention of its time.**

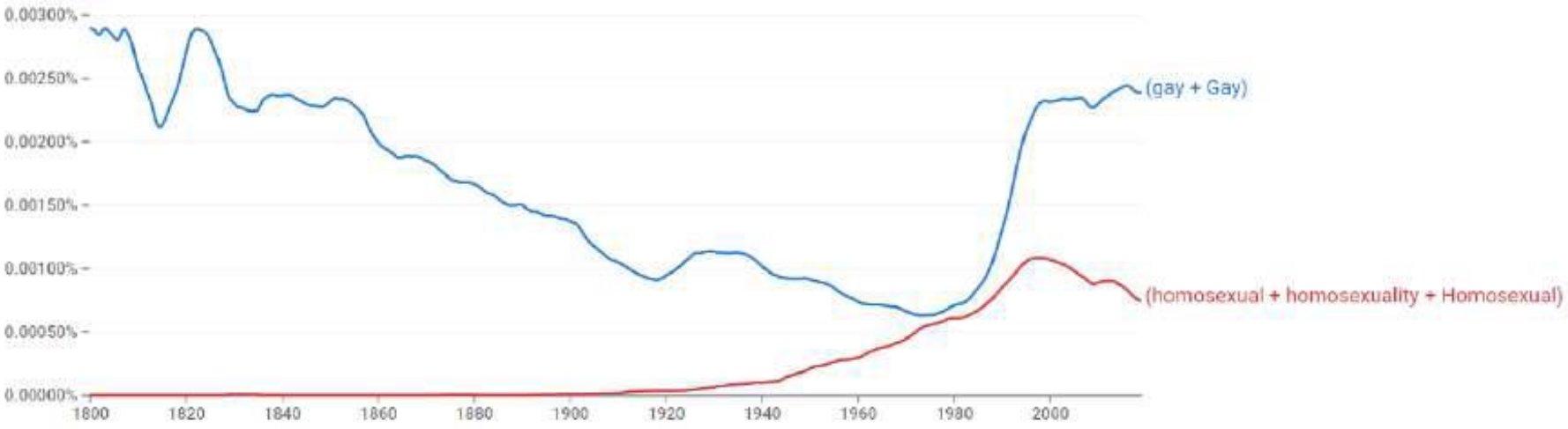
- You may choose affirmative or negative
- Prepare an introductory speech of two min.
  - After introductory speeches participants may oppose (for one min. each)
  - Then audience may oppose one side or the other
- Total should take about ten minutes

Examples (from 2021)

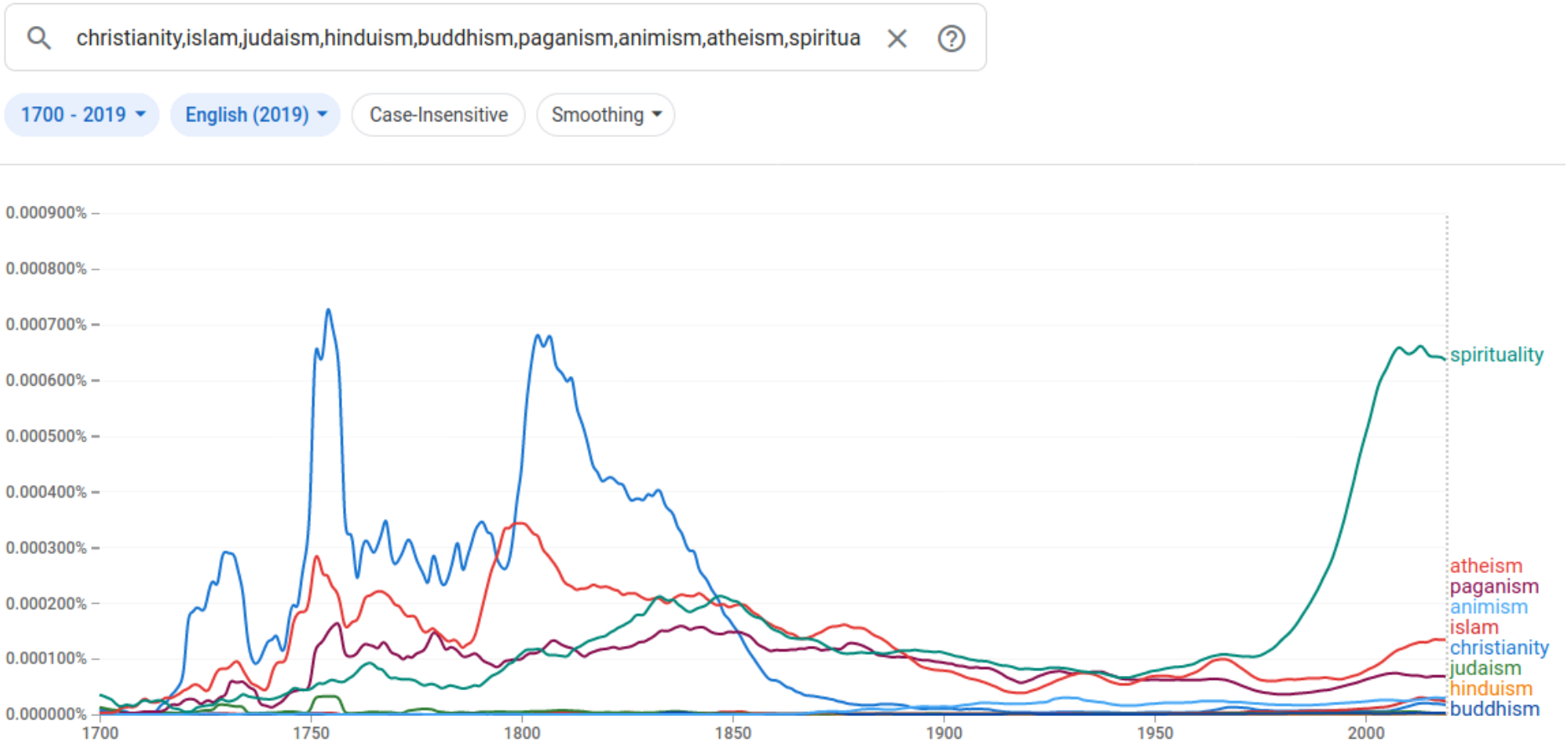
- Flu



Homosexual vs *gay*



Religious discourse



reich ‘rich’ vs Reich ‘empire’