# Notes on text mining 17th and 18th century English-language texts

Ryan Heuser, University of Cambridge
18 May 2020
Euro News Project

slides: ryanheuser.org/assets/slides/euronews2020.pdf
twitter: @quadrismegistus

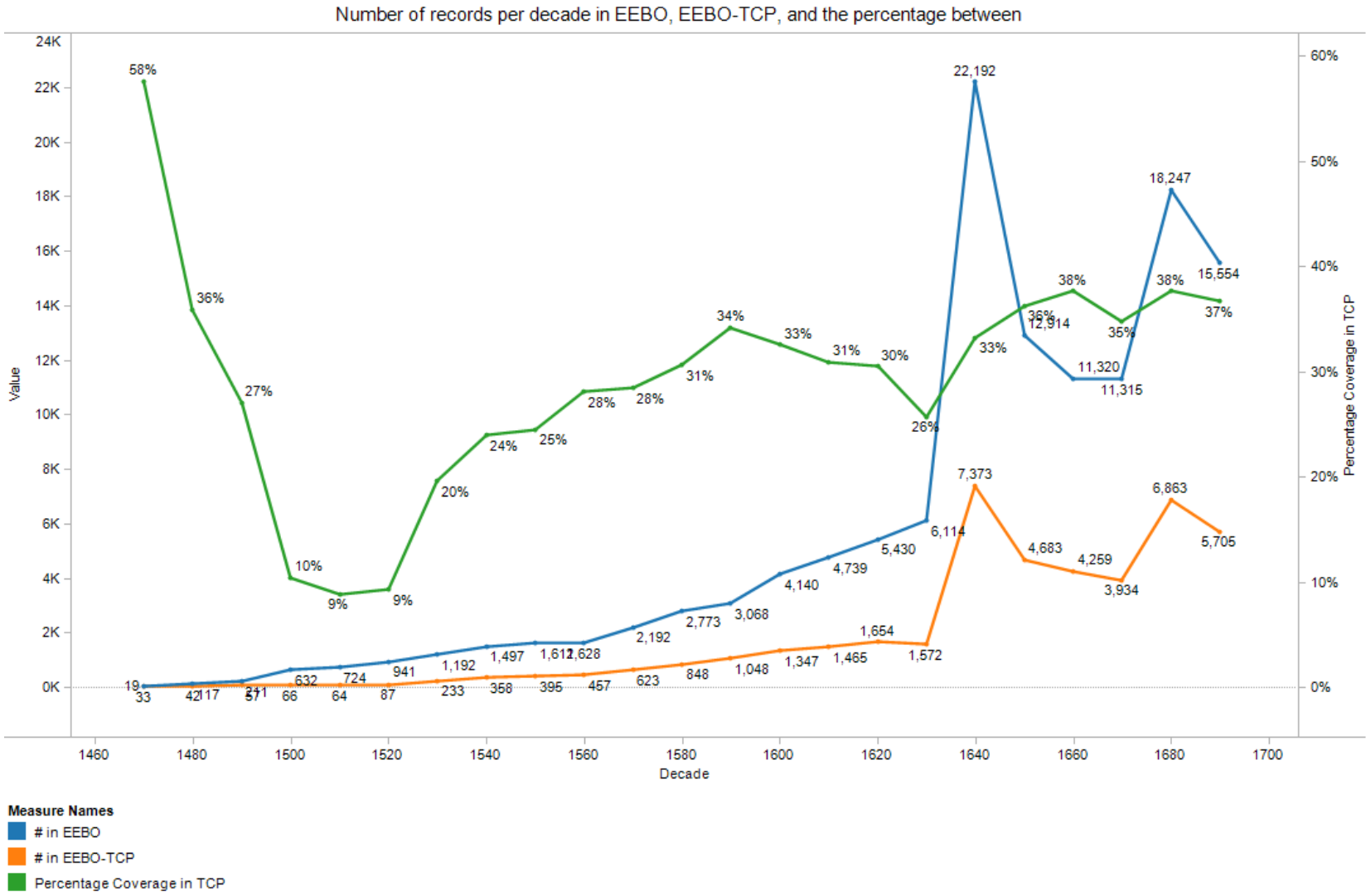
# Pre-18th century Corpora (in English)

- Early English Books Online (EEBO)
  - Approx. 128,000 printed texts
  - Sub-collections:
    - Short Title Catalogue, 1475-1640 (Pollard & Redgrave)
    - Short Title Catalogue, 1641-1700 (Wing)
    - Thomas Tracts
      - "Nearly everything" (approx. 80%) published between 1640-1660, including more than 400 periodicals as well as speeches, tracts, pamphlets, news reports, etc
    - Tract Supplement
      - "Scrapbooks" or tract volumes of broadsides, pamphlets, letters, ballads, etc.
  - Produced by ProQuest for library subscription

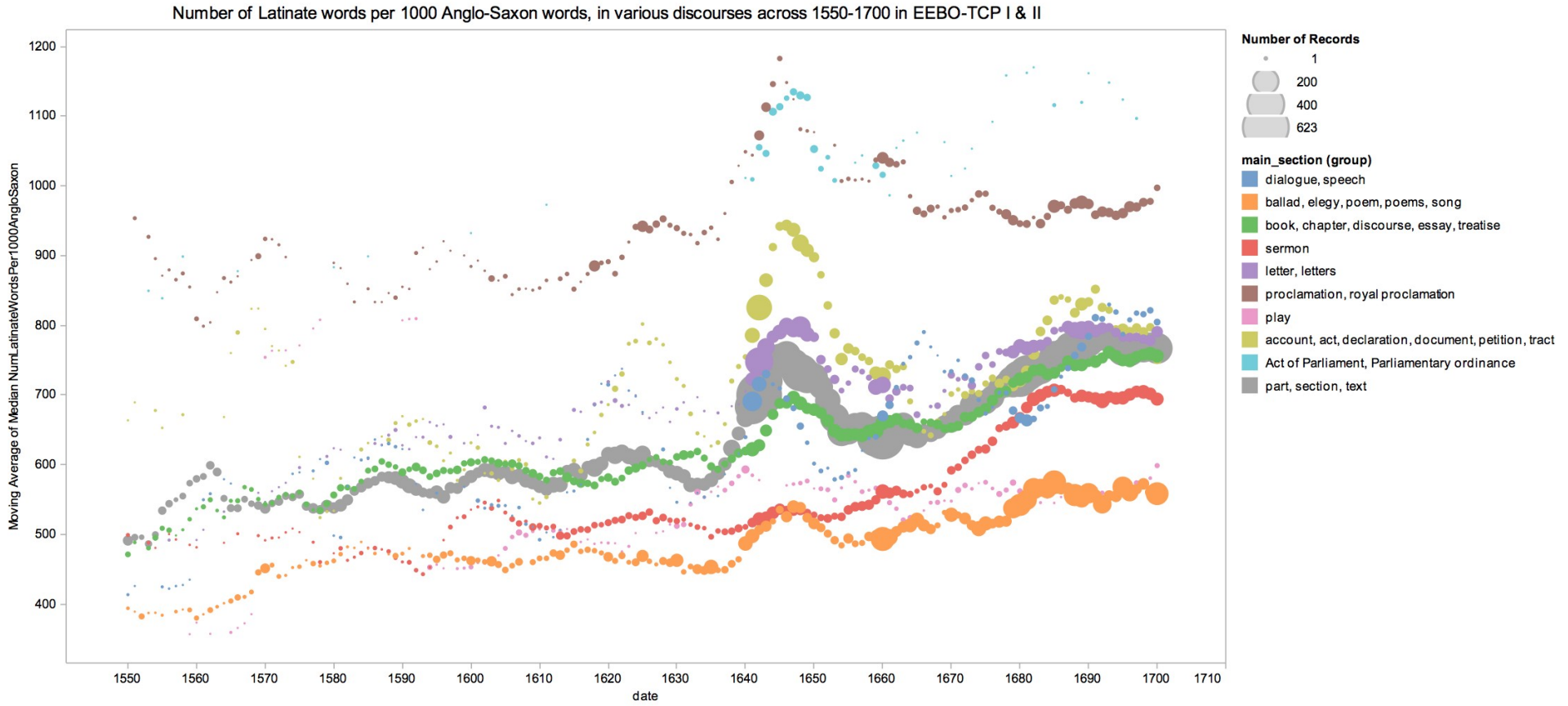
# Pre-18th century Corpora (in English)

- Early English Books Online – Text Creation Partnership (EEBO-TCP)
  - Sample of EEBO
  - Approx. 45,000 printed texts (~35% of EEBO)
  - All *hand-typed* (arranged by the Text Creation Partnership), with accurate rendering of spelling variants and encoded in XML
  - Texts selected if they or their author is the *New Cambridge Bibliography of English Literature* (NCBEL)
  - Metadata available:
    - Full title; Author name, date of birth, date of death
    - Which and how many libraries hold a copy of the text
    - Annotated title page, showing date of publication, publication place, publisher, any other information
    - Information embedded in document:
      - Section type of the document (“title page”, “to the reader”, “poem”, “proclamation”, “letter”, etc.)
      - Type of text (line of verse, paragraph, item in a list)
      - Page number

**Coverage of EEBO by EEBO-TCP**

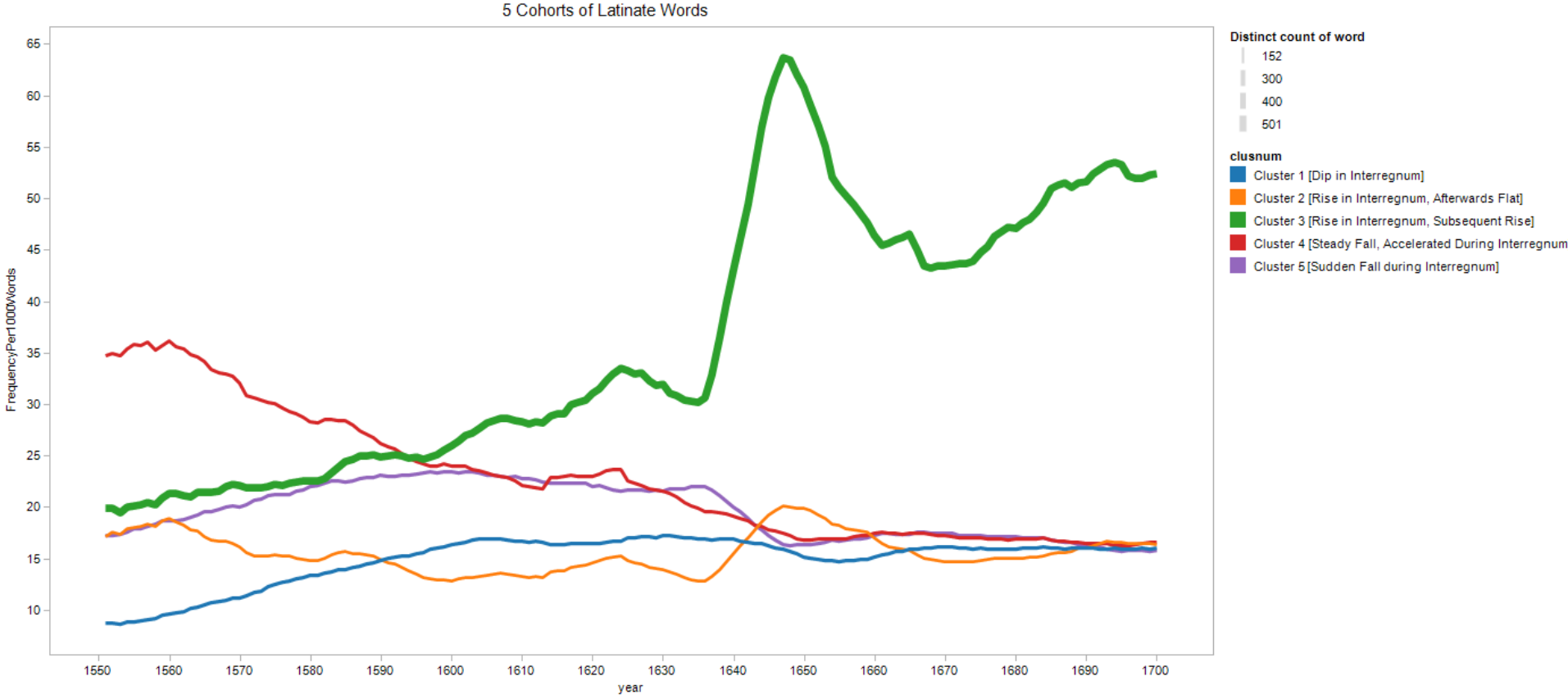


# Linguistic changes in EEBO-TCP, 1550-1700




The plot of Moving Average of Median NumLatinateWordsPer1000AngloSaxon for date. Color shows details about main_section (group). Size shows sum of Number of Records. The data is filtered on Exclusions (date,main_section), which specifies a set. The view is filtered on date and main_section (group). The date filter ranges from 1550 to 1700. The main_section (group) filter has multiple members selected.

# “Cohorts” of words which follow a similar historical trajectory in their frequency of use



"Cohorts" of words which follow a similar historical trajectory in their frequency of use

| Cluster | # of words | % of all Latinate occurrences | Words |
|---|---|---|---|
| Cluster 3 (Green) [Rise in Interregnum, Subsequent Rise] | 501 | 33% | very page sir persons power parliament reason present spirit act religion printed several state general royal case command justice due believe public appear duty divine concerning just conscience sense government pay nation honourable army return account **liberty** hereby consider captain particular whereas spiritual pleased acts letter obedience already christians private |
| Cluster 4 (Red) [Steady Fall] | 182 | 21% | majesty cause faith called continued neither common realm certain matter whereof ma contrary pass subjects pleasure save used reign ill princes pain **virtue** wicked force sort cum noble whereby doctrine purpose prayer vain suffer counsel sum doubt learned faithful grant seem wherefore plain quod lawful perfect example item except |
| Cluster 5 (Purple) [Sudden Fall during Interregnum] | 191 | 18% | like because use poor honour **prince** est pro country wherein diverse passed number joy age face praise sighed ut estate enter favour point strange si comfort servants danger stay labour gain saviour isle honest grief matters mere change regard died pains voice content delight confess parish hour fame increase |
| Cluster 2 (Orange) [Rise in Interregnum, Afterwards Flat] | 152 | 14% | people order person per money jesus desire charge authority judge sure able war receive christian enemies **officers** received necessary proclamation aforesaid ac appointed ministers scripture chief rule please occasion declare letters marry natural office effect promise cry minister committed scriptures continue delivered fine par execution piece defence declared papists trouble |
| Cluster 1 (Blue) [Dip in Interregnum] | 193 | 14% | found city nature second glory prey once judgement excellent whatsoever company want course happy subject respect soldiers enemy ancient get especially eternal sufficient opinion salvation gracious air privy **sacred** prayers thence presence commanded carry whence easily secondly measure pardon tune foreign built ordinary benefit weak ease pity sorts round quite |

Table 2: Five Latinate cohorts, and their fifty most frequent constituent words.

# 18th century Corpora (in English)

- Eighteenth Century Collections Online
  - Approx. 200,000 printed texts
    - “contains every significant English-language and foreign-language title printed in the United Kingdom between the years 1701 and 1800”
  - Across a range of genres, divided into 7 sub-collections:
    - History and Geography
    - Social Science and Fine Arts
    - Medicine
    - Science and Technology
    - Literature and Language
    - Religion and Philosophy
    - Law and General Reference
  - Once again, fairly poor OCR/text digitization quality
  - Produced by Gale

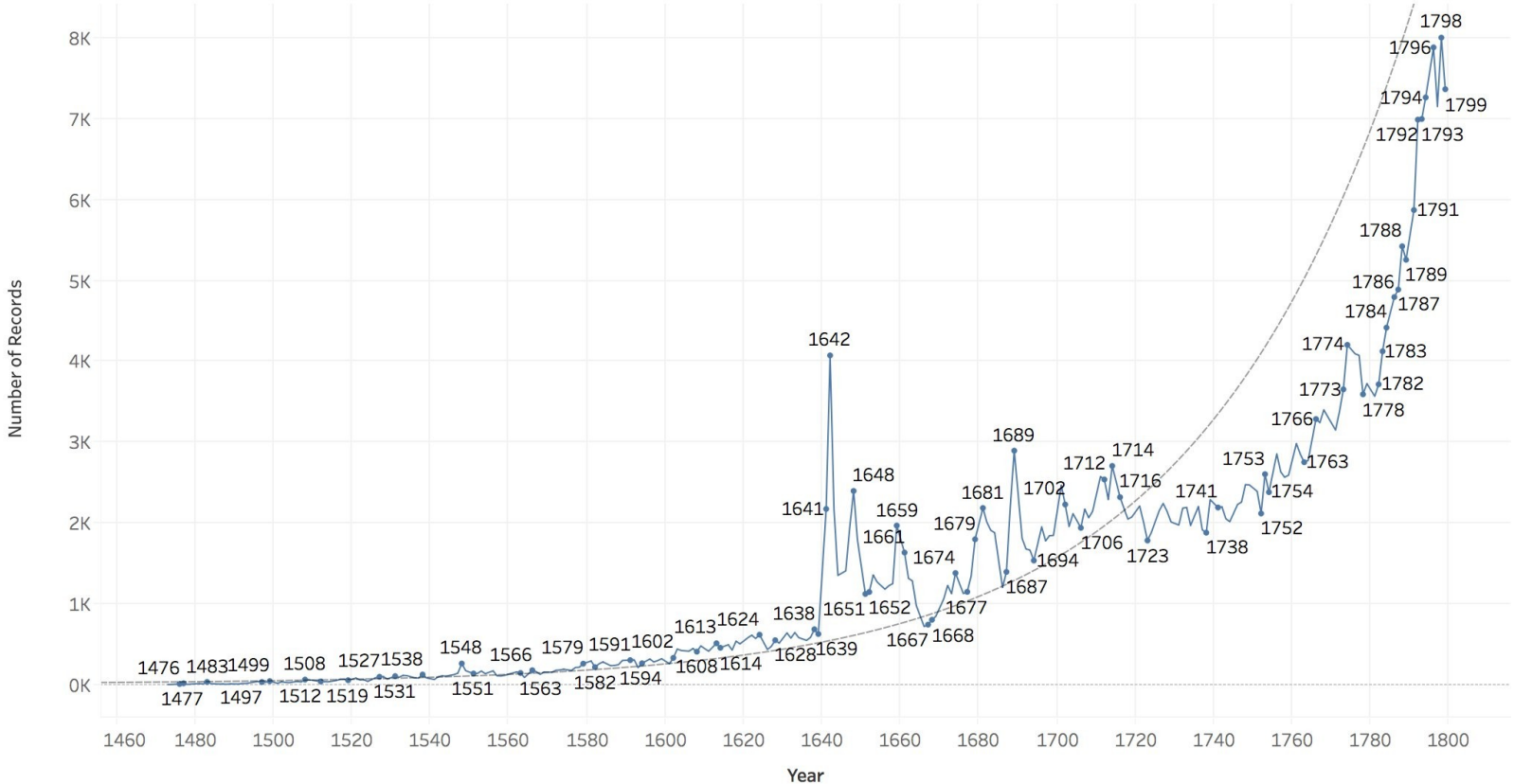
# 18th century Corpora (in English)

- Eighteenth Century Collections Online – Text Creation Partnership (ECCO-TCP)
  - 2,473 texts selected from EEBO (criteria unclear)
  - Hand-typed, so perfect digitized text quality
  - Encoded in TEI XML
  - Same metadata as ECCO

# Pre-1800 Bibliography

- English Short Title Catalogue
  - “a vast database designed to include a bibliographic record of every surviving copy of letterpress produced in Great Britain or any of its dependencies, in any language, worldwide, from 1473-1800”
  - Approx. 480,000 titles
  - No text: just metadata
  - But records keyed to the text file IDs in ECCO (and I believe EEBO as well)
  - Produced by the British Library

# Exponential rise of British publishing market, 1473-1800

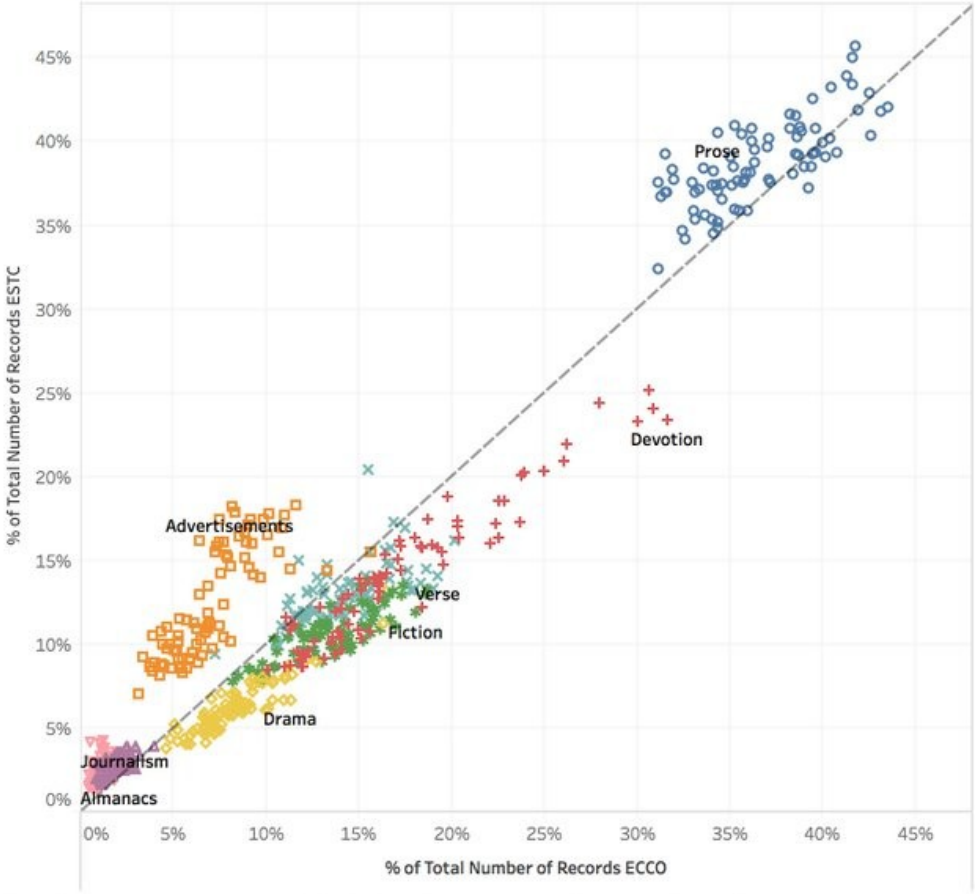



Number of books published per year since the beginning of the British printing press to 1800. The rise fits an exponential curve, with an R^2 value of 0.92. The deviations from this trend are significant and interesting: most notably, the explosion of pamphlets at the outset of the English Civil War in the early 1640s and during the Exclusion Crisis of the 1680s; and the recession of the publishing market in the first half of the eighteenth-century, due to the advent of government-regulated copyright in 1710 and the bursting of the South Sea Bubble in the 1720s. Data is taken from the English Short Title Catalogue, which is a "a vast database designed to include a bibliographic record, with holdings, of every surviving copy of letterpress produced in Great Britain or any of its dependencies, in any language, worldwide, from 1473-1800" (estc.ucr.edu). Alan Veylit has already extensively studied these publishing market trends in the ESTC (estc.ucr.edu/ ESTCStatistics.html).

# Corpus (ECCO)
# vs.
# Bibliography (ESTC)

**Q:** How representative of the authoritative 18C bibliography (ESTC) is the authoritative 18C text-database (ECCO)?

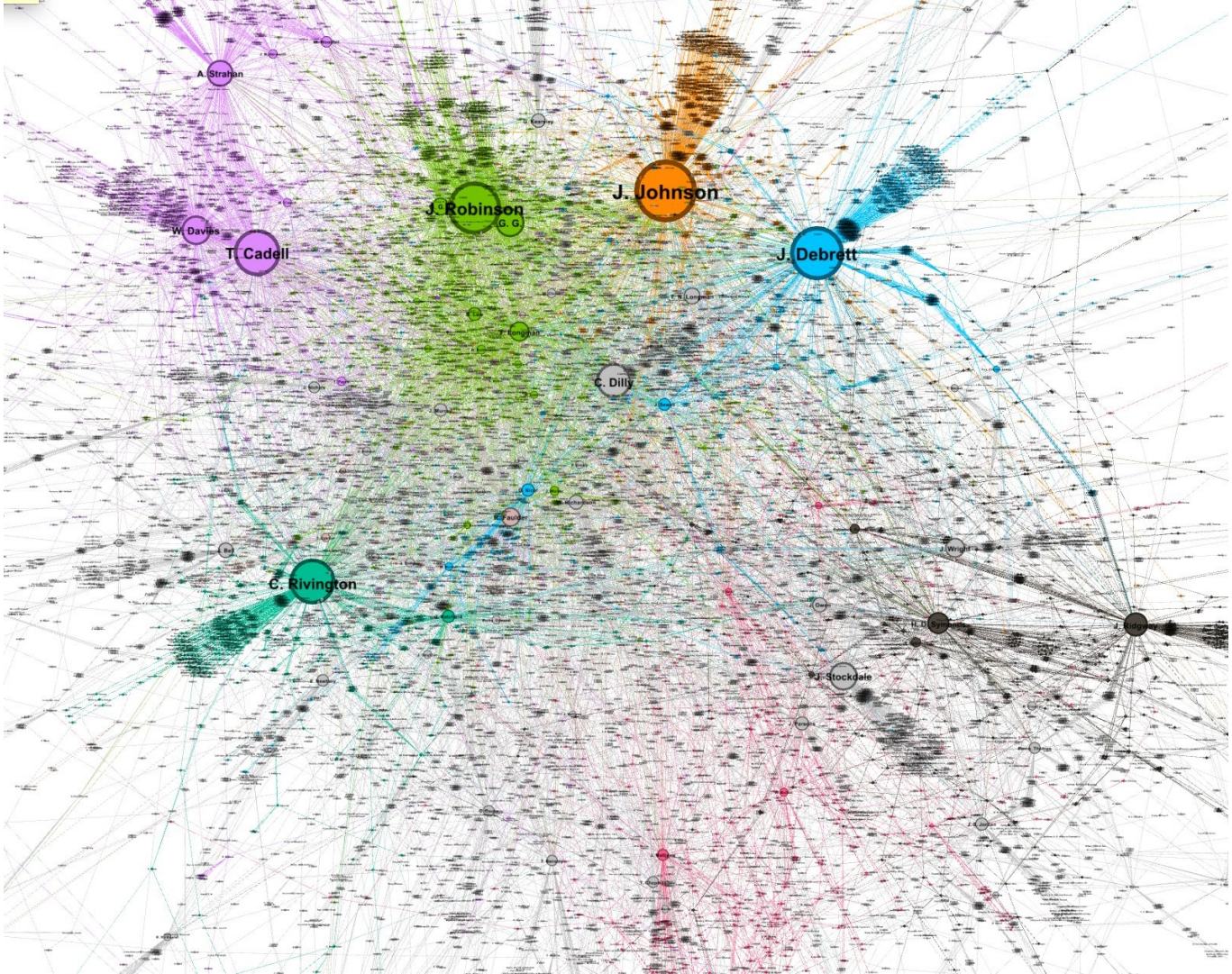
**A:** ECCO overrepresents "literary" genres (Drama, Fiction, Verse, and Sermons/Devotional literature) and underrepresents "non-literary" genres (Almanacs, Periodicals/Journalism, Advertisements, and Pamphlets/Miscellaneous Prose).




*The 45-degree line represents an idealized "balanced" representation: the relative density of a given genre in ECCO*

**Mapping London's print market**

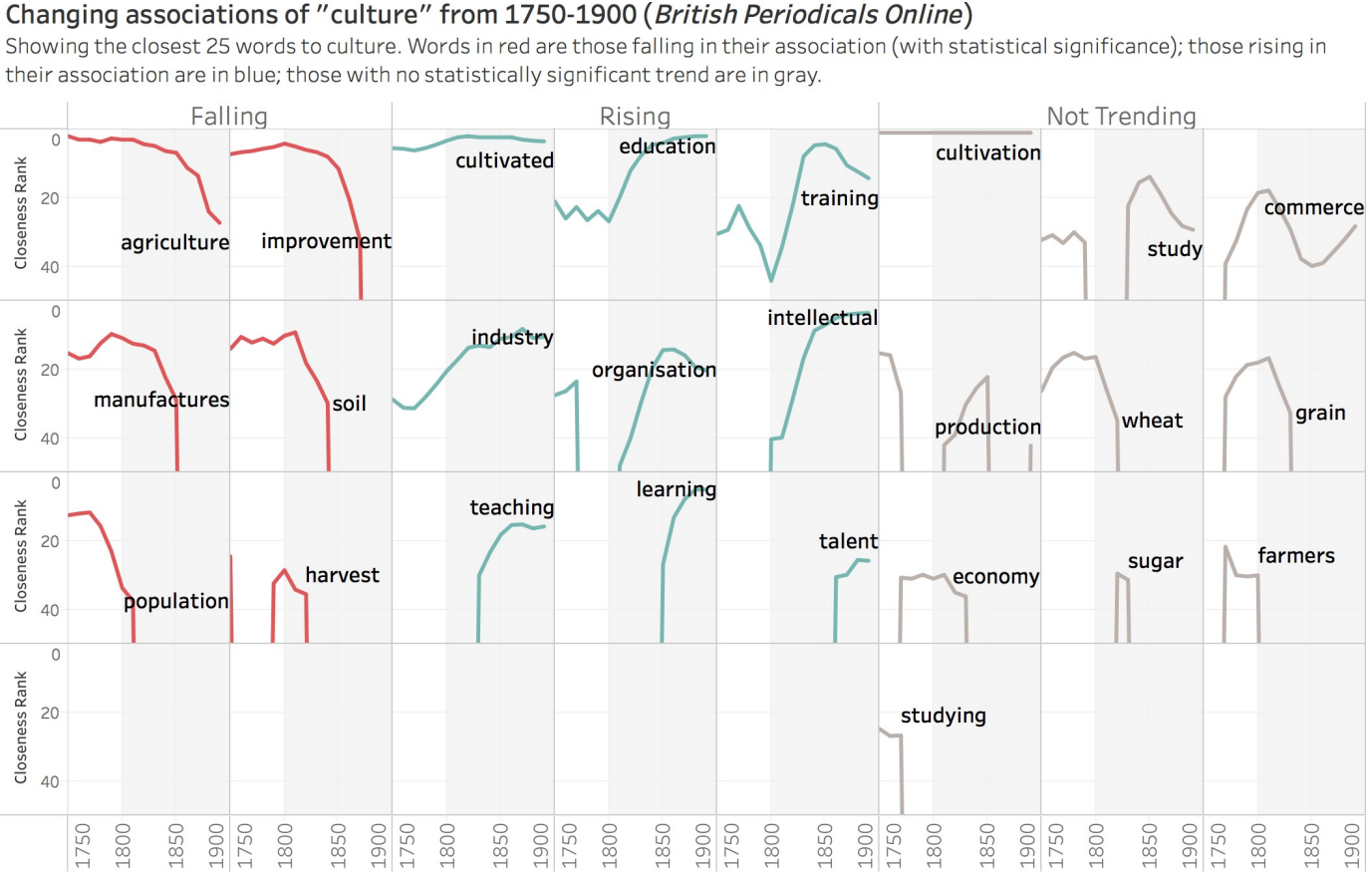
Network of texts and their publishers in the 1790s (ECCO data)



# Periodicals Corpora

- British Periodicals Online (1680-1900)
  - About ~3 milion periodical articles published in Britain between 1680 and 1900
  - About 4.4 billion words
  - OCR accuracy improves over time, but begins with poor quality
  - Produced by ProQuest

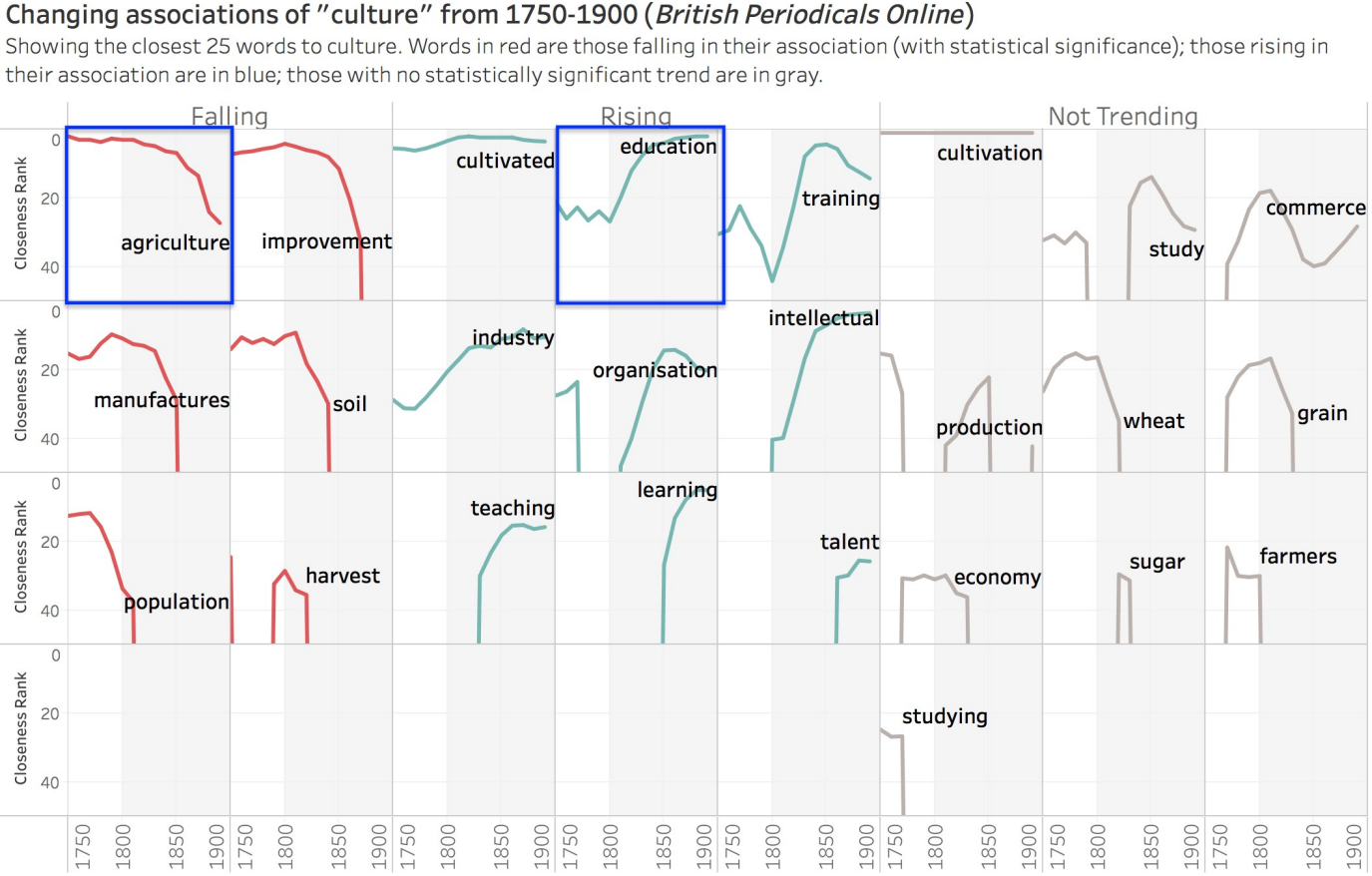
# Testing hypothesized semantic shifts




## Culture

"Before this period [late 18C], it had meant, primarily, the 'tending of natural growth', and then, by analogy, a process of human training. But this latter use, which had usually been a culture of something, was changed, in the nineteenth century, to culture as such, a thing in itself. It came to mean, first, 'a general state or habit of the mind', having close relations with the idea of human perfection. Second, it came to mean 'the general state of intellectual development, in a society as a whole'. Third, it came to mean 'the general body of the arts'. Fourth, later in the century, it came to mean 'a whole way of life, material, intellectual and

# Testing hypothesized semantic shifts




## Culture

"Before this period [late 18C], it had meant, primarily, the 'tending of natural growth', and then, by analogy, a process of human training. But this latter use, which had usually been a culture of something, was changed, in the nineteenth century, to culture as such, a thing in itself. It came to mean, first, 'a general state or habit of the mind', having close relations with the idea of human perfection. Second, it came to mean 'the general state of intellectual development, in a society as a whole'. Third, it came to mean 'the general body of the arts'. Fourth, later in the century, it came to mean 'a whole way of life, material, intellectual and

# Other historical corpora

**In English**

- EEBO (Early English Books Online), 1473-1799
- Google Books / HathiTrust, ~1700-2000
- Corpus of Historical American English, 1810-2010
  - https://www.english-corpora.org/coha/
- Chadwyck-Healey Literature Online, ~1600-2000
  - Poetry, Fiction, and Drama

**Other languages**

- Google Books, ~1700-2000
  - French, English, Spanish, Russian, Italian, Hebrew, Chinese (simplified)
- French: ARTFL
- German: Deutches Text Archiv