Sunoikisis Digital Classics Spring 2025

Analysing and Visualising Text

Kaspar Beelen (University of London)
Megan Bushnell (Oxford Text Archive)

Outline of course (delete?)

- Introduction of course convenors?
- Intro to Digital Text Analysis (20-30 minutes)
  - ‘Distant reading’ /‘Big data’ [KB] (5 mins)
  - Goals of digital text analysis [KB] (5 mins)
  - Corpus linguistics [MB] (3 mins)
    - History [MB] (2 mins)
  - Historical data (& warnings) [MB] (3 mins)
  - Applications (one slide) [MB] (2 mins)
- Case studies (20-30 minutes)
  - Greek example/Latin example [GB] (5 minutes)
  - Latin example/medieval Scots example [MB] (5-10 minutes)
  - Short Parliamentary Example [KB] (5-10 minutes)
- Intro to Voyant Tools (20-30 minutes) [KB]
  - Showcase different panels (15 minutes)
  - Intro to exercise (5mins)

Outline

- Introduction
- [Digital] Text Analysis
- Case studies
- Voyant Tools

[Digital] Text Analysis

Information Overflow?

Question: how to “read” a million (or more) books?

Our ability to read and process information is limited

- We are not immortal, alas…
- We need sleep and our attention span/memory has limits

Information Overflow?

- As “humanists/classicists” we/you are trained to contextualize and interpret a carefully selected set of sources
- Works well on small amount of materials, but does not scale up easily

Question rephrased: can we say something meaningful based on “big historical data?

“Distant Reading”

“[a] little pact with the devil: we know how to read texts, now let’s learn how not to read them.” (Moretti)

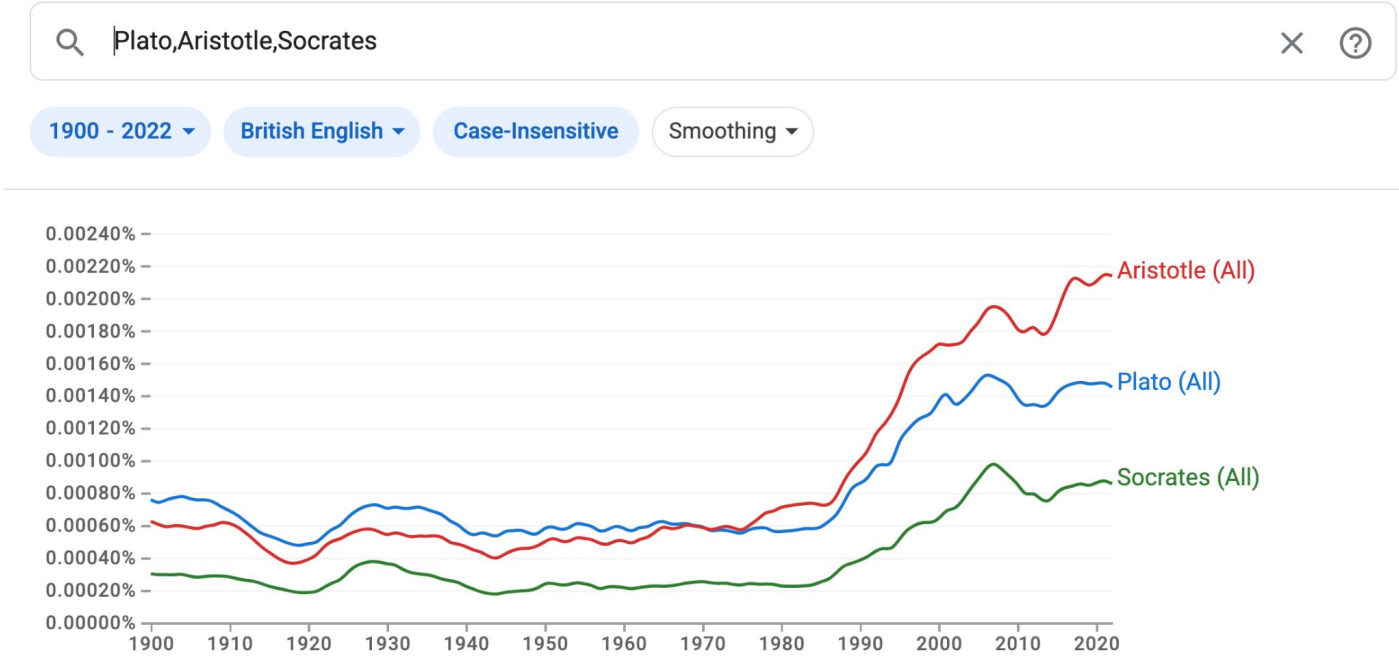
- **observe** text “en masse”.
- **foreground structures** that only emerge when analysing texts at **scale**.
- **dissolve texts**,tearing them apart, transforming them to more abstract units.
- **quantify** text using automated methods for counting.

“Distant Reading”

The practice of distant reading is therefore sometimes referred to as the ‘abattoir of literature’.

An Example of “Distant Reading”

Does the N-Gram corpus show an increasing popularity of classical authors?



https://books.google.com/ngrams/

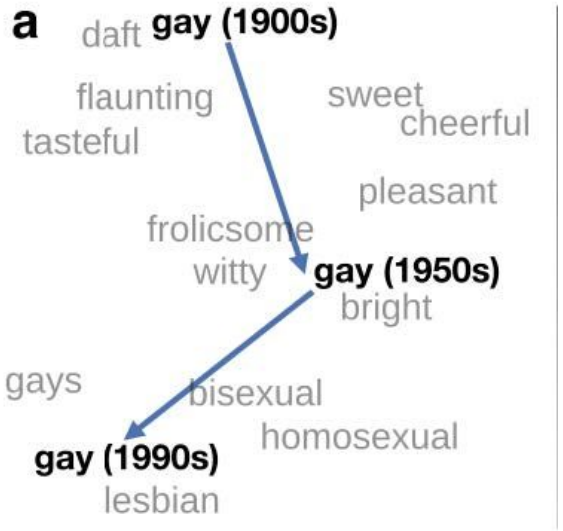
“Reading without reading”

Can we use textual evidence via distant reading (i.e.without *actually* reading these books?)

- **Observe en masse**: obtain a snapshot based on millions of books
- **Dissolve** texts through quantification: data in the backend consists of word counts (“bag-of-words representation”)
- **Foreground structure**: relation between data (frequency of tokens) and metadata (language and time)

(Digital) Texts-as-Evidence? Critical Questions.

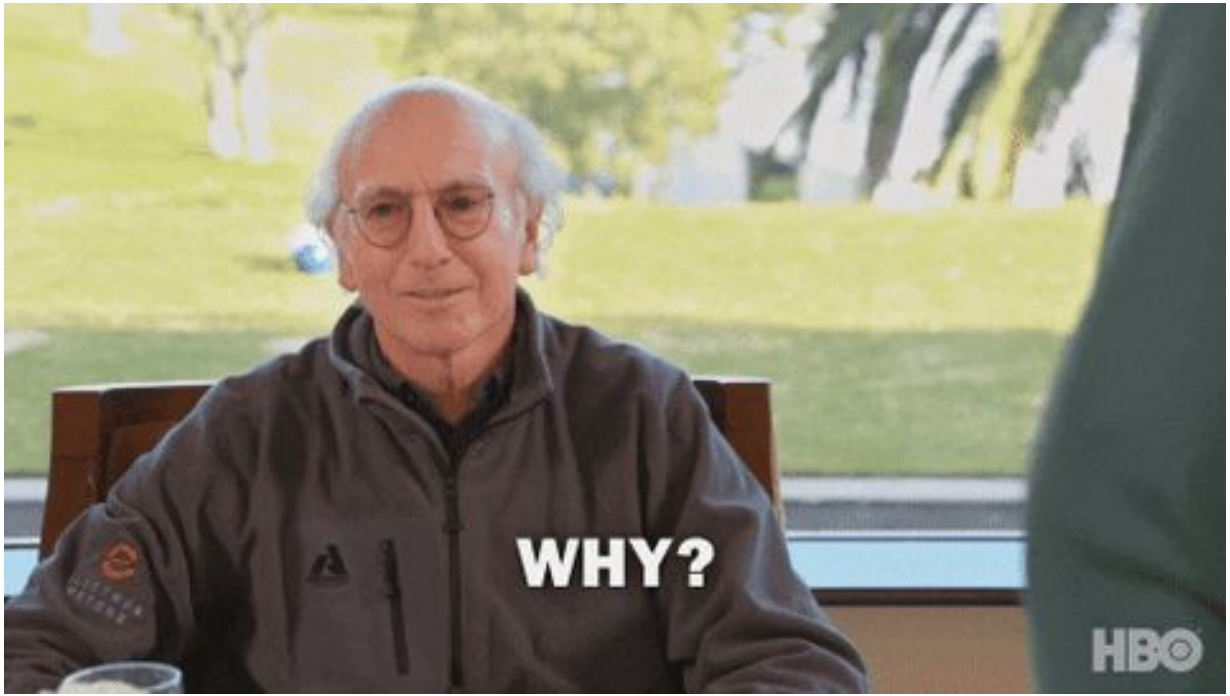
- **But what does it mean?** Does the N-Gram corpus suggest classical authors are becoming increasingly popular?
  - Is the data **representative**: what does the data “stand for” as a collection?
  - What do the **words** mean? Are they **ambiguous**?
  - Are the words **stable** in **meaning**? Is the data **stable** in its **composition**?
  - What does **frequency** mean?



Source: https://arxiv.org/pdf/1605.09096

WHY?? Aims of Digital Text Analysis

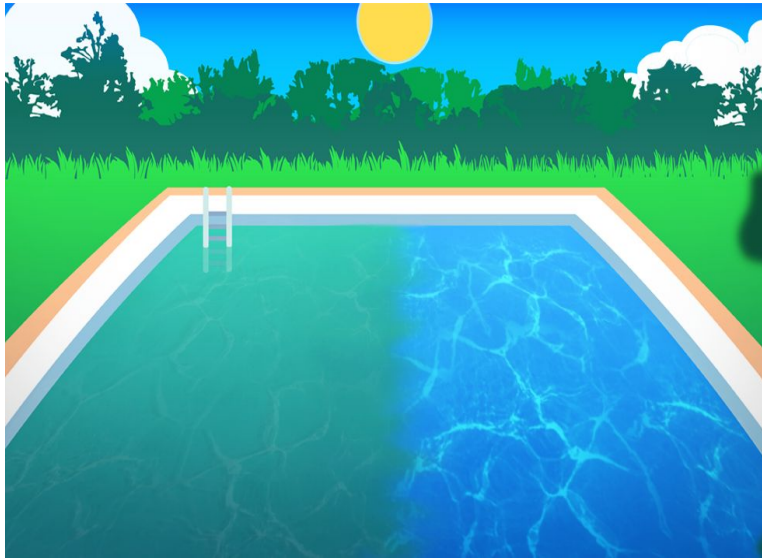
- Discovery
- Measurement
- Interpretation



Discovery and Exploration

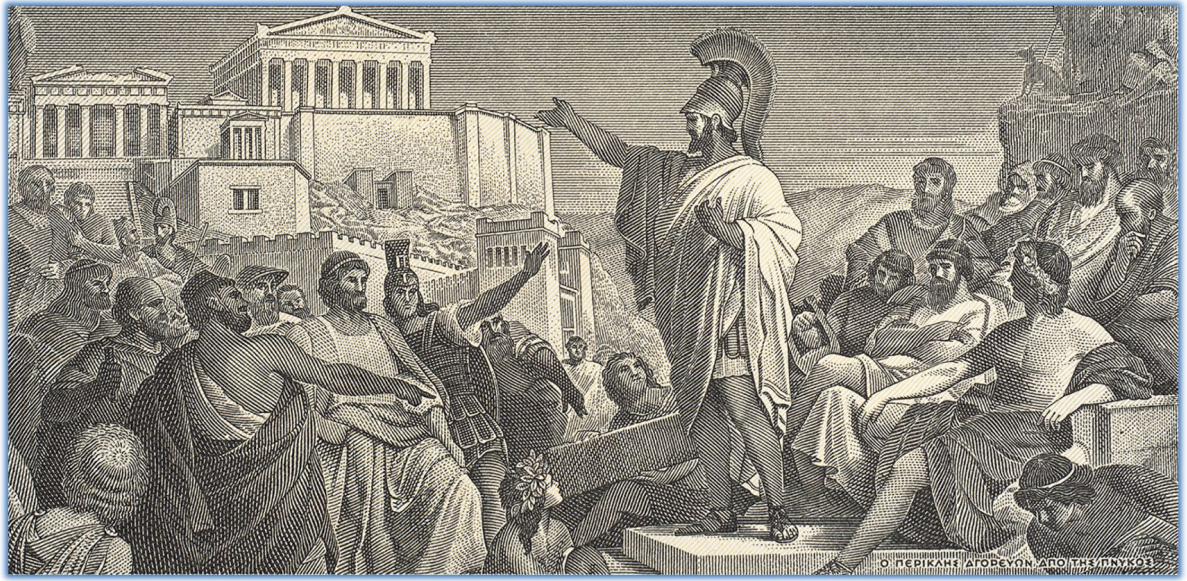
When you use texts as evidence, you first need to get to know your data!

- Collections we work with are often “**opaque pools of information**”
- **Bias** and **representativeness** are pervasive problems… not just for digital humanities
- Gaps and silences



Large Language Models based on historical text could offer informative tools for behavioral science

Michael E. W. Varnum, Nicolas Baumard, M



By training them on various types of historical texts and records, LLMs could help scholars better grasp the mentality of ancient peoples during major historical events. Here, Pericles gives a famous speech around 430 BC, a year after the start of the Peloponnesian War. Image credit: Shutterstock/vkilikov.

Techniques for Discovery: **Search**

- Digital search has opened up radically new ways of **exploring** large corpora of text. With just a few **keystrokes** we can navigate our way to specific **content** in large text collections. However …
- **Keyword search is a form of digital text analysis**
  - A lot of NLP happens in the background: lemmatization, removal of stopwords, weighing words, ranking documents
  - In what order are results sorted?
  - What am I actually searching for?
  - Is my (re)search reproducible? Can I document my process?

Techniques for Discovery: **Word Clouds**

Word cloud based on Mary Shelley's "Frankenstein"

Measurement

As humanists, we are (usually) comfortable with the concept of **discovery**

The idea of **measurement** is maybe unfamiliar. Can you measure literary texts the same way you record, for example, temperature or fungus growth?

What is measurement: we quantify certain aspects of a text as a proxy for some concept.

- “happy words” to measure emotion
- “Sentence length” for readability/difficulty
- “Stop words” for authorship

Example: “The Loudness in the Novel”

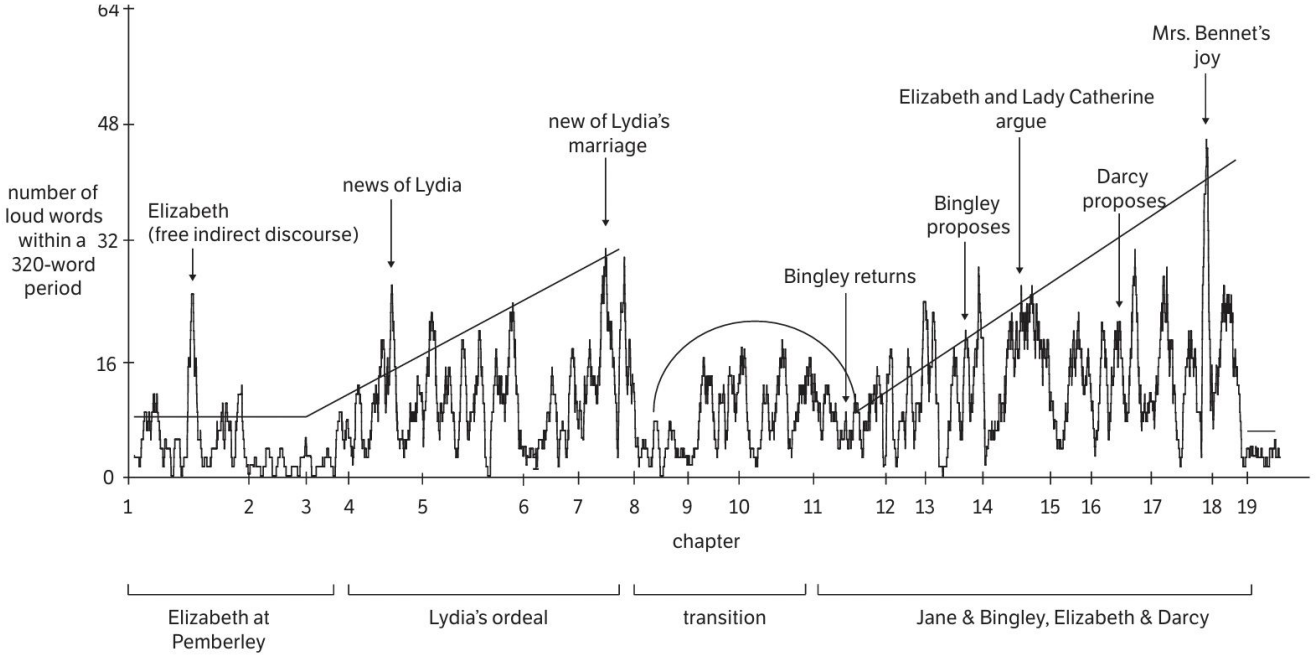


Figure 6: Loudness in *Pride and Prejudice*, Volume 3

Source: https://litlab.stanford.edu/projects/loudness/

Conceptualisation and “Construct Validity”

Conceptualisation is a critical element of digital text analysis. If we want to study a phenomenon with computational means, we must:

1. Clearly articulate the concept of interest and;

2. Explain how we intend to measure it

Semantic errors are prevalent in computational text analysis.“If our machines don’t measure what we tell them to, our analysis and findings fall apart quickly.” (Kaspar Beelen, 2024)

- Do you really measure happiness when counting happy words?

Lastly: Interpretation is King

From text numbers… and back again?

- Build a narrative through discovery and measurement
- This is often an iterative and circular process

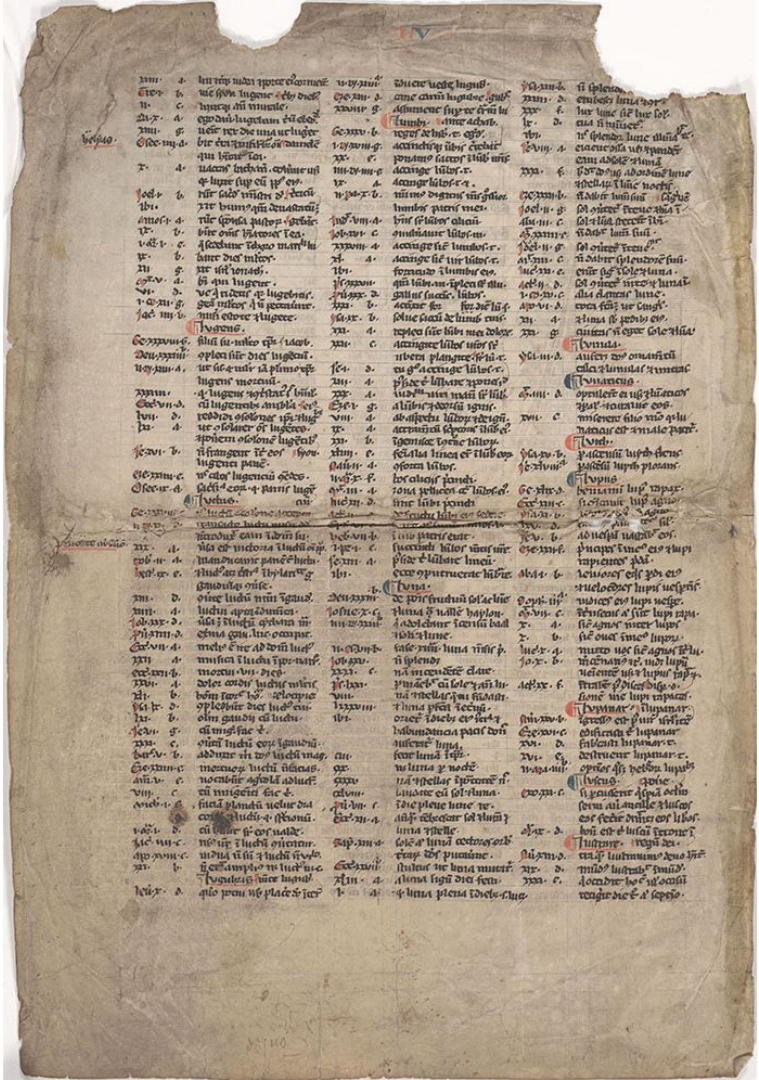
History of Digital Text Analysis

Corpus linguistics

- The study of language in a written context - specifically a body of works, or corpus.
- Concordancer:
  - Key word in context - or KWIC.
  - Collocates - words with a strong tendency to co-occur.
  - Keywords - words that are distinct in one text or texts compared to another/others.
  - N-grams - lexicalised collocates.
  - AntConc - https://www.laurenceanthony.net/software/antconc/.
  - LancsBox - https://lancsbox.lancs.ac.uk/.
- Annotation
  - Lemmatisation
  - Part-of-speech tagging
  - Semantic tagging
  - And much more!

Not as new as we think?

- Enumerative bibliography - assembling national bibliographies.
- Gematria - transforming words into numbers.
- *Sortes Vergilianae, sortes Biblicae* - fortune telling.
- Cryptography and word frequency.
- Commentaries and glossaries.
- Bible concordances and *distinctiones*.



[Bible Concordance: Luc - Lus]. *Free Library of Philadelphia*: Philadelphia, PA. https://libwww.freelibrary.org/digital/item/19483. (accessed Nov 4, 2024)

Historical data health warnings!

- Data may require more processing before analysis.
  - Lack of standardisation.
  - Lack of resources to help you process your data.
- Data may suffer especially from a lack of context.
  - Material.
  - Cultural.
  - Historical.
  - Linguistic.
  - Several layers of abstraction.
- Data will likely be smaller, and less representative.

Applications

- Linguistic analysis
  - Lexicography
  - Language acquisition
  - Translation
  - Discourse analysis
  - Language variation and change
- Literary analysis
  - Stylometry
- Historical analysis
- Computer science
  - NLP
  - LLMs
  - AI

Case Studies

Case study 1: Claudius Aelianus' *De Natura Animalium*

- Claudius Aelianus in the Perseus Catalog: urn:cite:perseus:author.19
- *De Natura Animalium* in the Perseus Catalog: urn:cts:greekLit:tlg0545.tlg001
- *De Natura Animalium* in the PerseusDL: https://github.com/PerseusDL/canonical-greekLit/tree/master/data/tlg0545/tlg001
- Aelian's *NA* to download: 17 txt files in a ZIP archive -- in Ancient Greek; lemmatised forms -- single file, Greek dictionary headings.
- Voyant Tools visualisations to browse: Raw text (in 17 books); Lemmatised text (in single file)

See more in SunoikisisDC 2022:
https://github.com/SunoikisisDC/SunoikisisDC-2021-2022/wiki/SunoikisisDC-Summer-2022-Session-2

Case study 2: N-grams and Repetitions

‘… **Connubio iungam stabili**: **propriamque dicabo.** …’ (*Aen.* I.73)

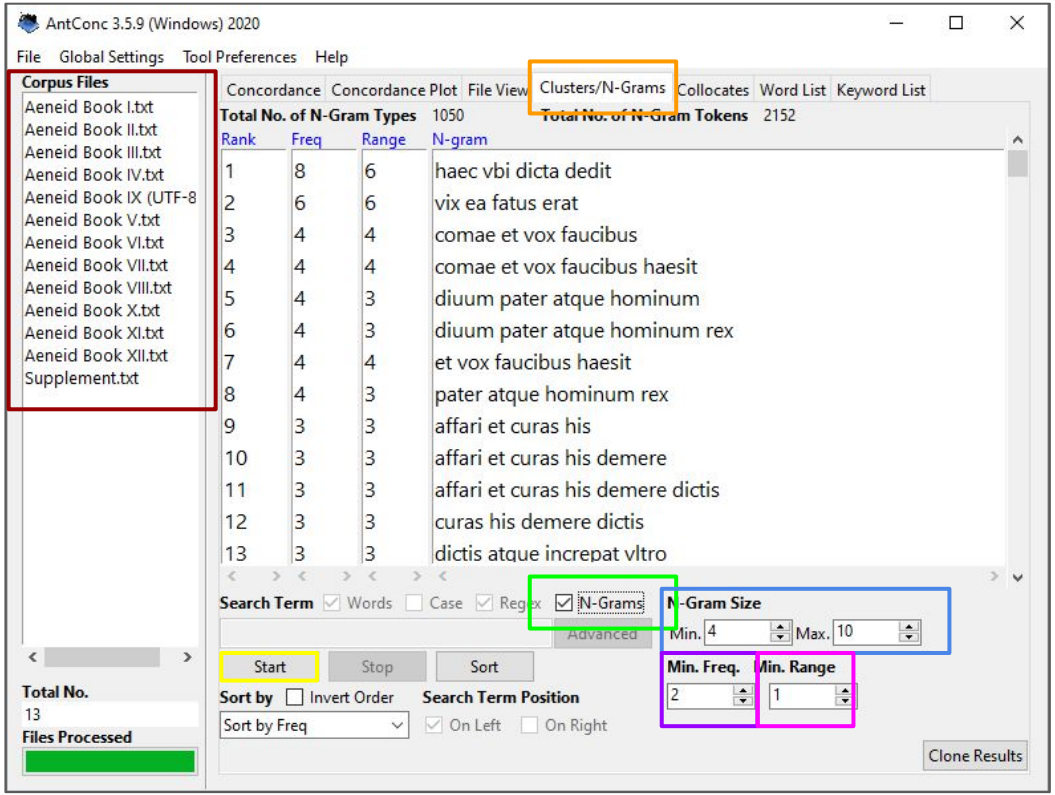
‘… I [Juno] would dedicate [Deiopea] as a steadfast wife and devoted spouse [to you, Aeolus]. …’ (my translation)

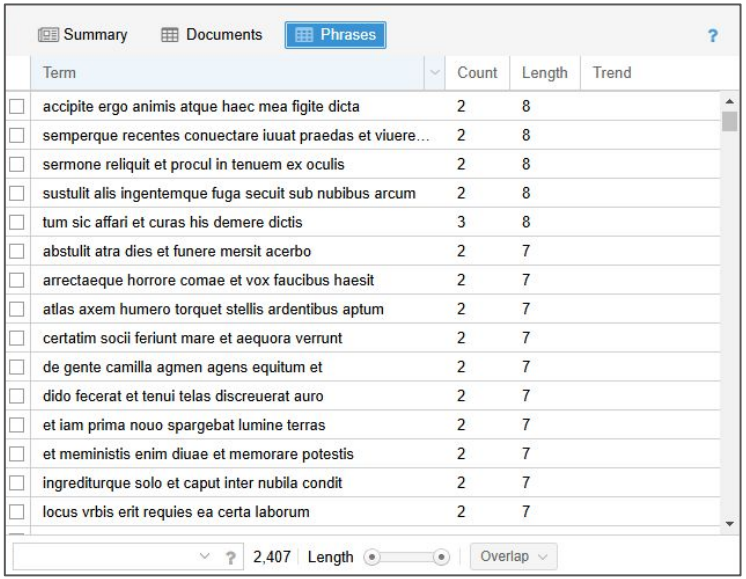
‘… **Connubio iungam stabili propriamque dicabo.** …’ (*Aen.* IV.126)

‘… I [Juno] would dedicate [Dido] as a steadfast wife and devoted spouse [to Aeneas]. …’ (my translation)

Landscape with the Union of Dido and Aeneas by Gaspard Dughet and Carlo Maratta, in *The National Gallery*, CC BY-NC-ND 4.0 (https://www.nationalgallery.org.uk/paintings/gaspard-dughet-and-carlo-maratta-landscape-with-the-union-of-dido-and-aeneas)

N-grams in AntConc and Voyant

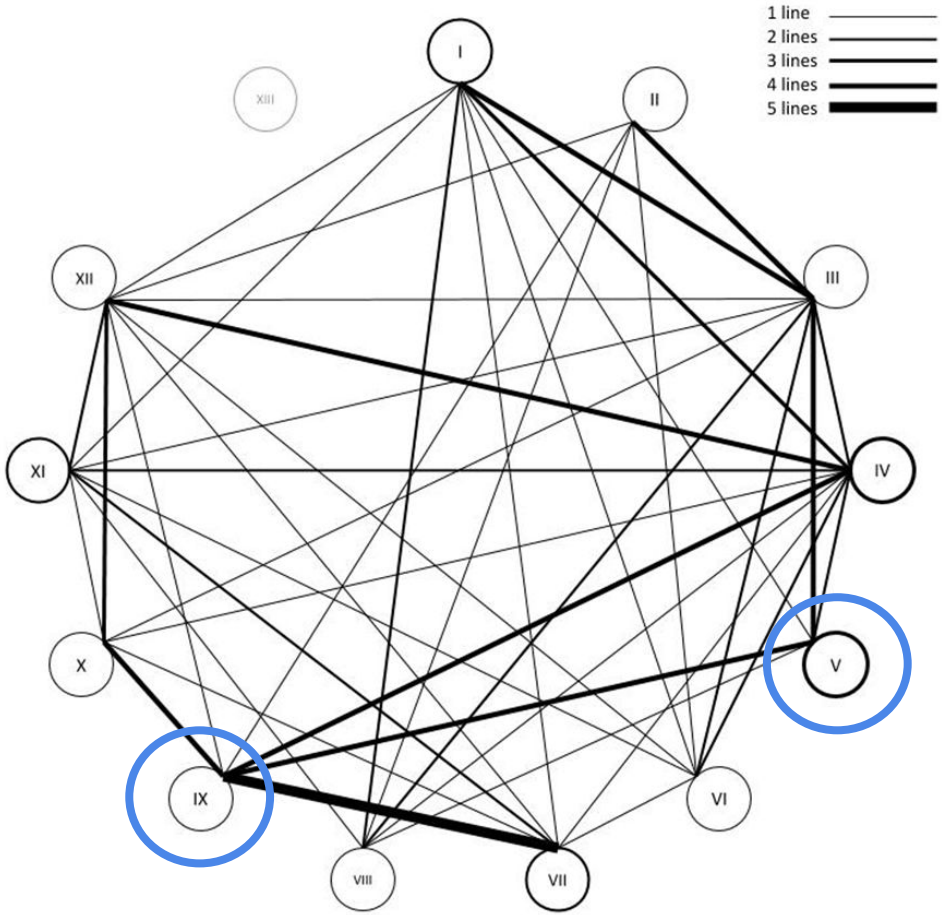




Repetitions in the *Aeneid*

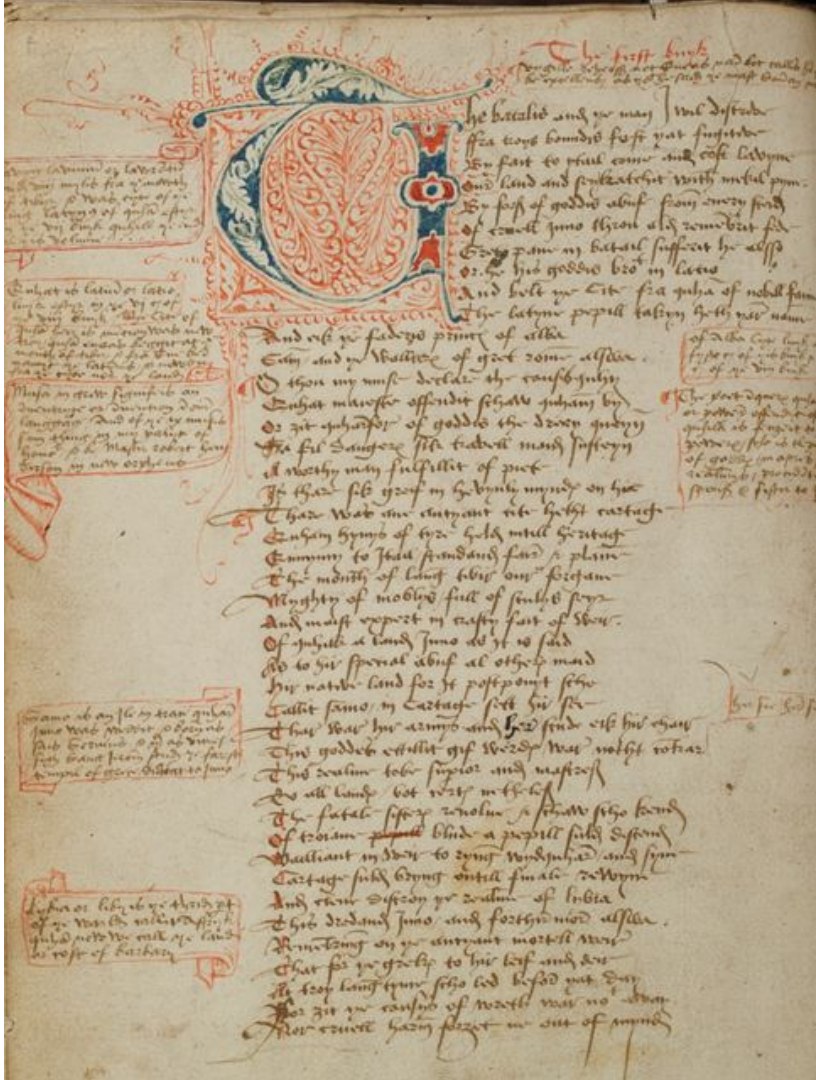
IV.272-88

Si te **nulla** mouet **tantarum gloria rerum**:
Nec super ipse tua moliris **laude laborem**: **(IV. 232-33)**
Ascanium surgentem: et spes haeredis iuli
Respice: cui regnum italiae romanaque tellus
Debentur. tali cyllenius ore locutus:
Mortales visus **medio sermone reliquit**:
Et procul in tenuem ex oculis evanuit auram. **(IX.657-58)**
At vero aeneas aspectu obmutuit amens:
Arrectaeque horrore comae: et vox faucibus haesit: **(XII.868)**
Ardet abire fuga: dulcesque relinquere terras.
Attonitus tanto monitu imperioque deorum:
Heu quid agat? quo nunc reginam ambire furentem
Audeat affatu: et quae prima exordia sumat:
Atque animum nunc huc celerem nunc diuidit illuc: **(VIII.20-21)**
In partesque rapit varias: perque omnia versat:
Haec alternanti potior sententia visa est:
Mnesthea: **sergestumque vocat**: **fortemque** cloanthum: **(XII.561)**



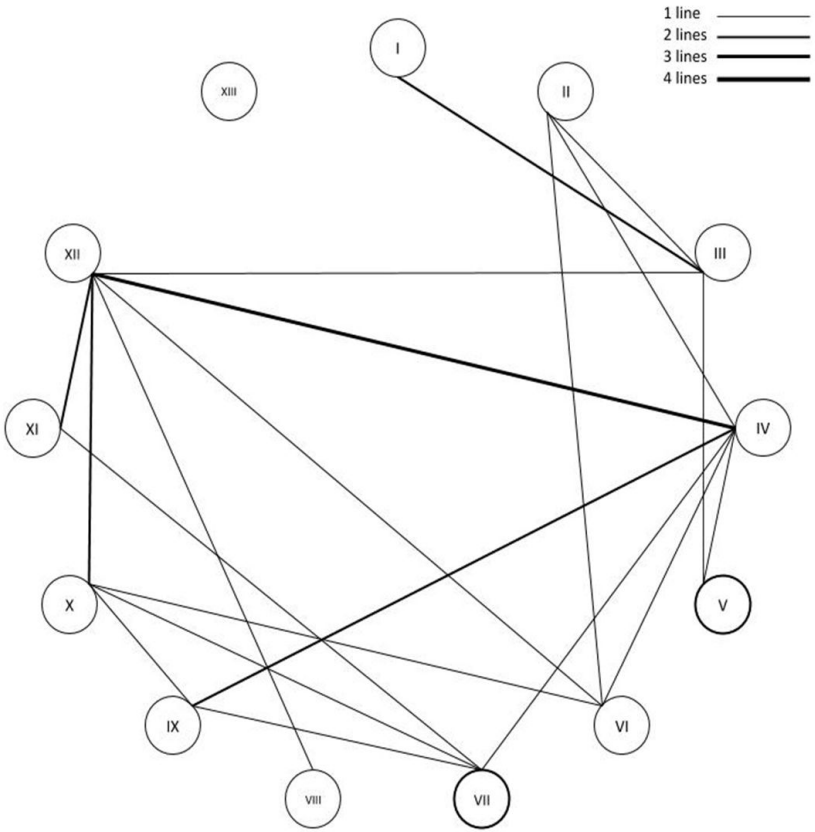
The *Eneados*

- Gavin Douglas (1474-1522)
  - Scottish cleric, nobleman, and poet.
- Eneados (1513)
  - First full translation of the *Aeneid*. Source text is Ascensius's (1501) edition (Bawcutt 1973).
- Includes extra material.
  - Prologues
  - Book XIII (translated from Maffeo Vegio's *Supplement*, 1428).
  - Twice as long as the original (21,047 vs. 9,867 lines).



Repetitions in the *Eneados*

- 257 repeated segments of at least 5 words.
- 54% are motivated by the source.
- Replicates 25% of line-length repetitions in the *Aeneid*.
- Ascensius/Servius (1501) only catch 6%.
- Dryden (1697) only catches 7%.
- Williams (1910) only catches 10%.
- 83% of Douglas's sourced repetitions have matching contexts.



Case Study 3: A Feminized Language of Democracy?

Gender and Politics in Westminster (1945-2013)

Digging into Data

A Feminized Language of Democracy?

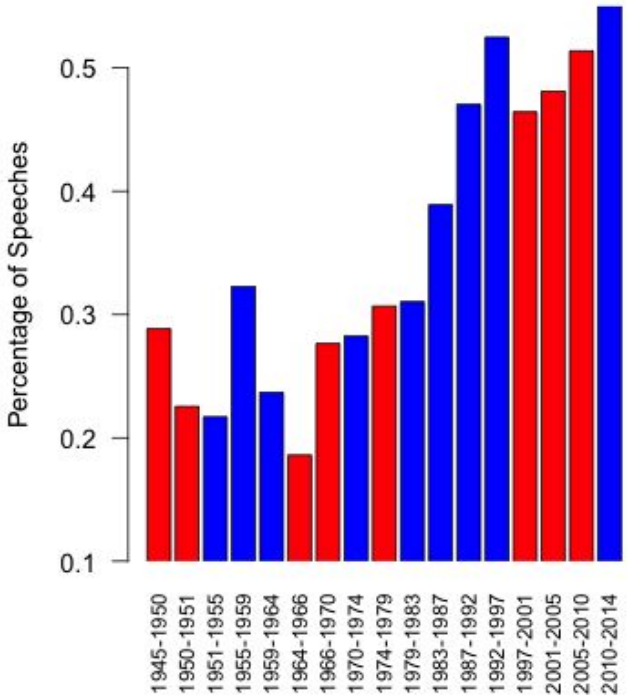
Research Questions:

- To what extent do female MPs act for women? Or change in terms of their legislative priorities?
- Has the increasing number of women MPs at Westminster influenced the “substantive” representation of women? (Politics of Presence)

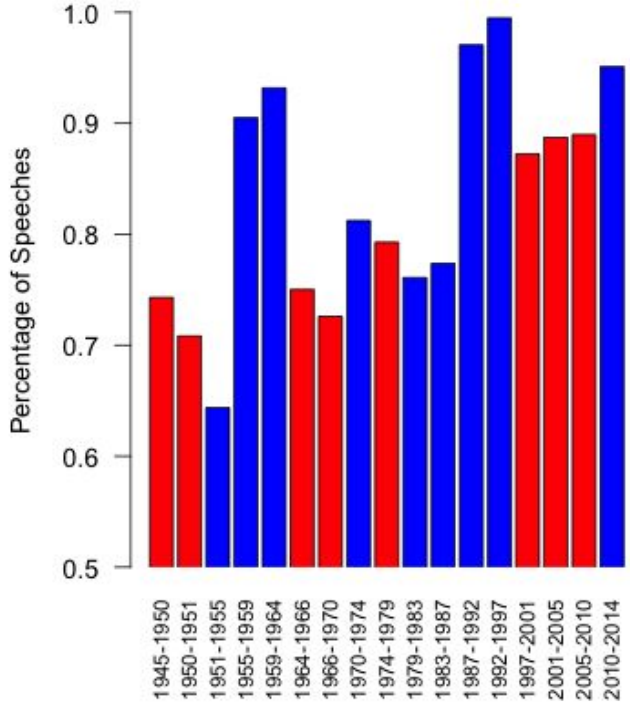
Method: measure the presence of gendered words such as ‘woman’, ‘mothers’ and study their relation to the MPs background

Case Study

Women [1]

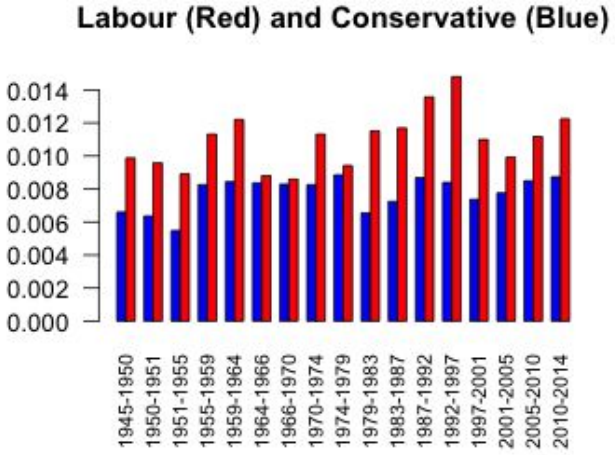


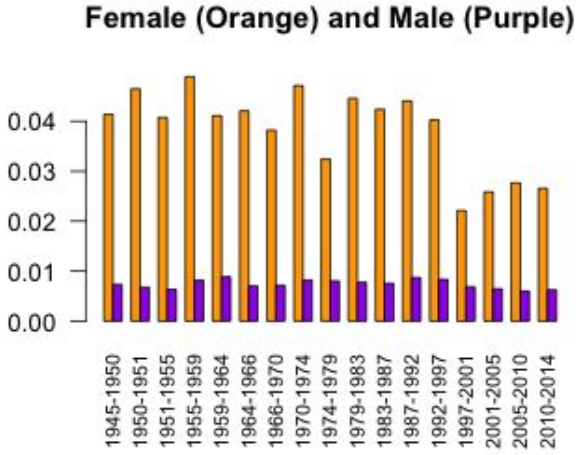
Women [2]

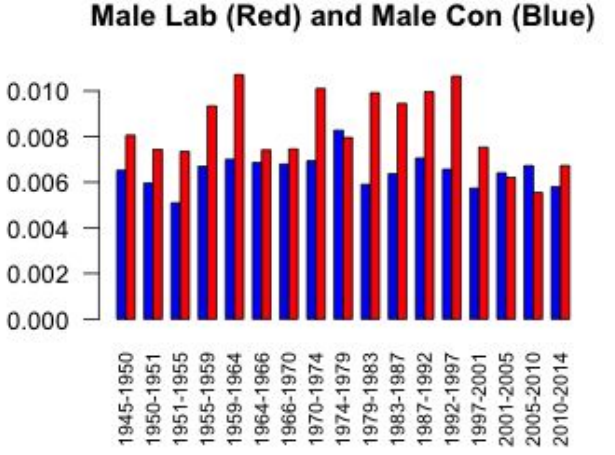


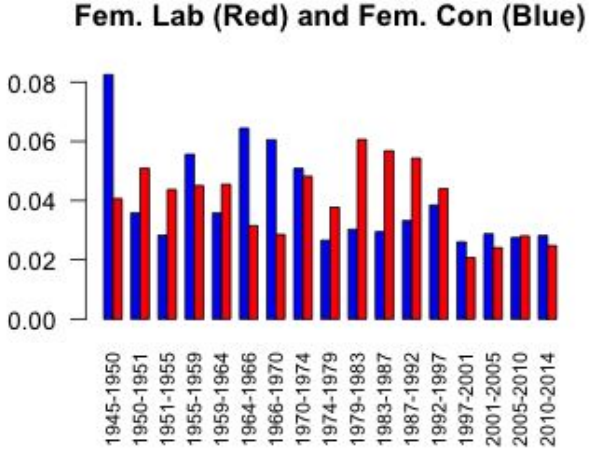
Blue = Conservative Majority/Coalition
Red = Labour Majority

Case Study

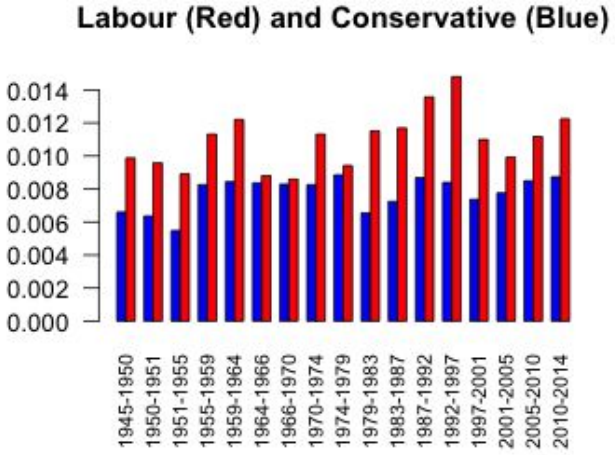

Labour (Red) and Conservative (Blue)
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
1945-1950
1950-1951
1951-1955
1955-1959
1959-1964
1964-1966
1966-1970
1970-1974
1974-1979
1979-1983
1983-1987
1987-1992
1992-1997
1997-2001
2001-2005
2005-2010
2010-2014

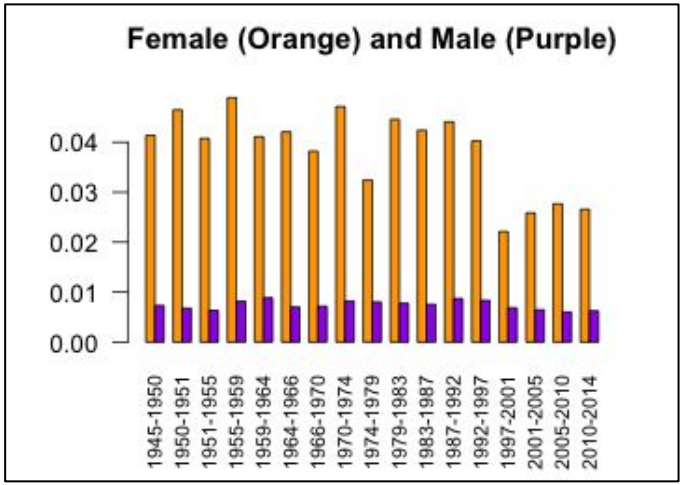

Female (Orange) and Male (Purple)
0.04
0.03
0.02
0.01
0.00
1945-1950
1950-1951
1951-1955
1955-1959
1959-1964
1964-1966
1966-1970
1970-1974
1974-1979
1979-1983
1983-1987
1987-1992
1992-1997
1997-2001
2001-2005
2005-2010
2010-2014

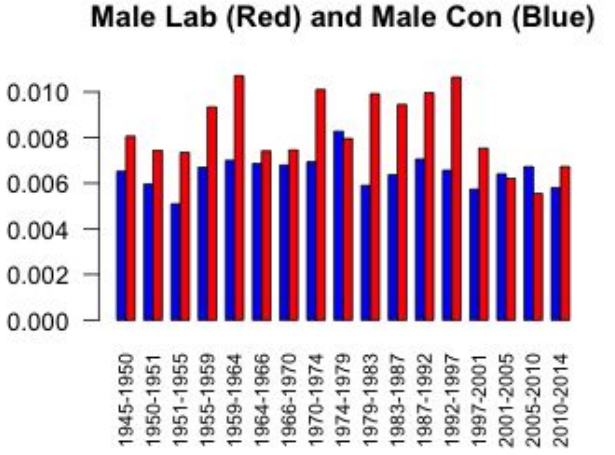

Male Lab (Red) and Male Con (Blue)
0.010
0.008
0.006
0.004
0.002
0.000
1945-1950
1950-1951
1951-1955
1955-1959
1959-1964
1964-1966
1966-1970
1970-1974
1974-1979
1979-1983
1983-1987
1987-1992
1992-1997
1997-2001
2001-2005
2005-2010
2010-2014

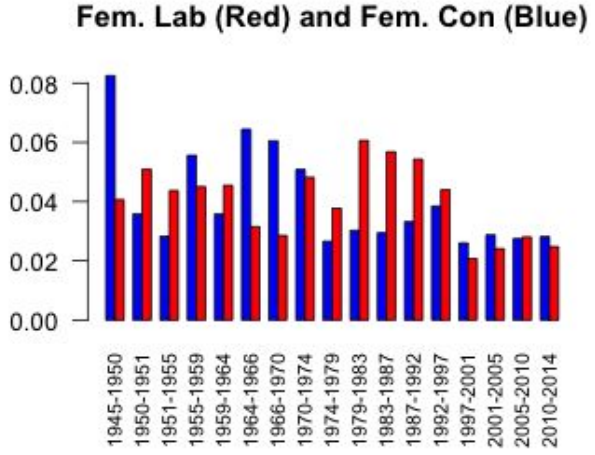

Fem. Lab (Red) and Fem. Con (Blue)
0.08
0.06
0.04
0.02
0.00
1945-1950
1950-1951
1951-1955
1955-1959
1959-1964
1964-1966
1966-1970
1970-1974
1974-1979
1979-1983
1983-1987
1987-1992
1992-1997
1997-2001
2001-2005
2005-2010
2010-2014

Case Study

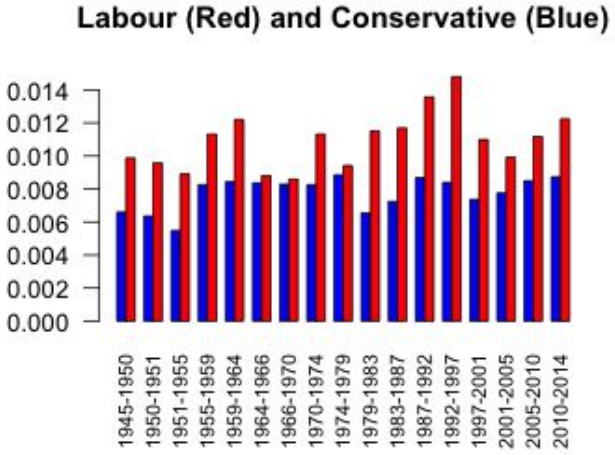

Labour (Red) and Conservative (Blue)
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
1945-1950
1950-1951
1951-1955
1955-1959
1959-1964
1964-1966
1966-1970
1970-1974
1974-1979
1979-1983
1983-1987
1987-1992
1992-1997
1997-2001
2001-2005
2005-2010
2010-2014

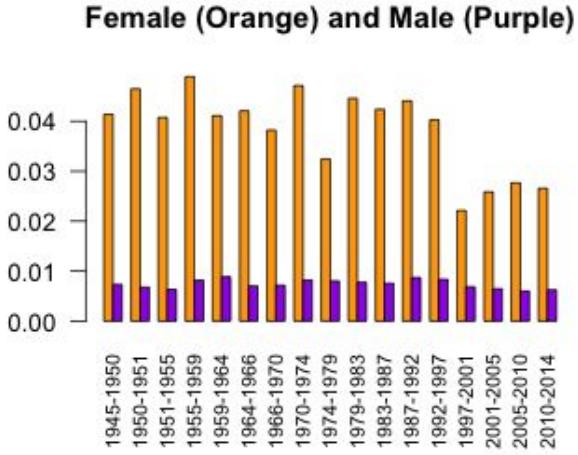

Female (Orange) and Male (Purple)
0.04
0.03
0.02
0.01
0.00
1945-1950
1950-1951
1951-1955
1955-1959
1959-1964
1964-1966
1966-1970
1970-1974
1974-1979
1979-1983
1983-1987
1987-1992
1992-1997
1997-2001
2001-2005
2005-2010
2010-2014

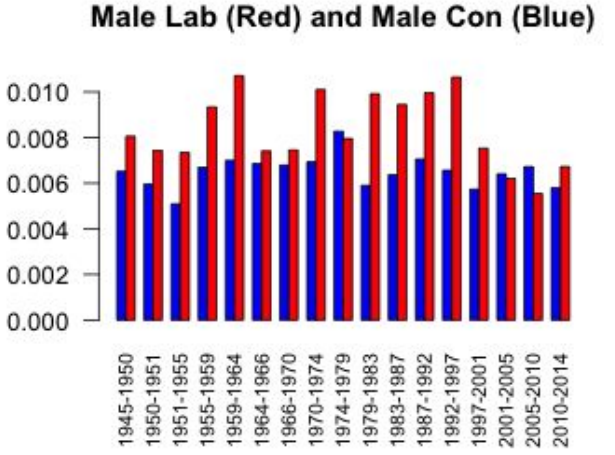

Male Lab (Red) and Male Con (Blue)
0.010
0.008
0.006
0.004
0.002
0.000
1945-1950
1950-1951
1951-1955
1955-1959
1959-1964
1964-1966
1966-1970
1970-1974
1974-1979
1979-1983
1983-1987
1987-1992
1992-1997
1997-2001
2001-2005
2005-2010
2010-2014

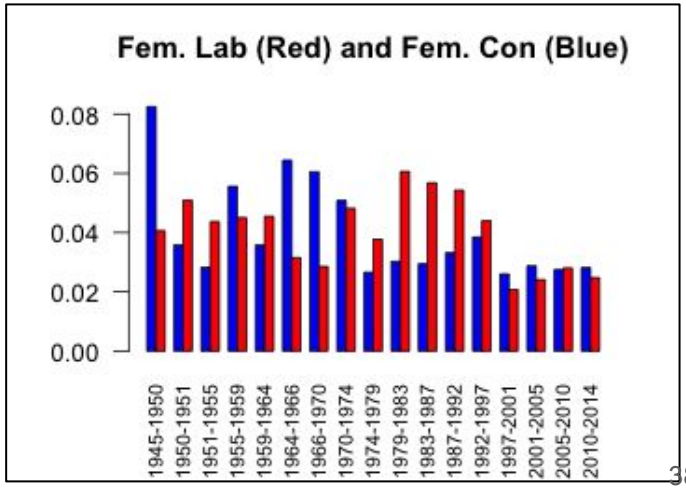

Fem. Lab (Red) and Fem. Con (Blue)
0.08
0.06
0.04
0.02
0.00
1945-1950
1950-1951
1951-1955
1955-1959
1959-1964
1964-1966
1966-1970
1970-1974
1974-1979
1979-1983
1983-1987
1987-1992
1992-1997
1997-2001
2001-2005
2005-2010
2010-2014

Case Study


Labour (Red) and Conservative (Blue)
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
1945-1950
1950-1951
1951-1955
1955-1959
1959-1964
1964-1966
1966-1970
1970-1974
1974-1979
1979-1983
1983-1987
1987-1992
1992-1997
1997-2001
2001-2005
2005-2010
2010-2014


Female (Orange) and Male (Purple)
0.04
0.03
0.02
0.01
0.00
1945-1950
1950-1951
1951-1955
1955-1959
1959-1964
1964-1966
1966-1970
1970-1974
1974-1979
1979-1983
1983-1987
1987-1992
1992-1997
1997-2001
2001-2005
2005-2010
2010-2014

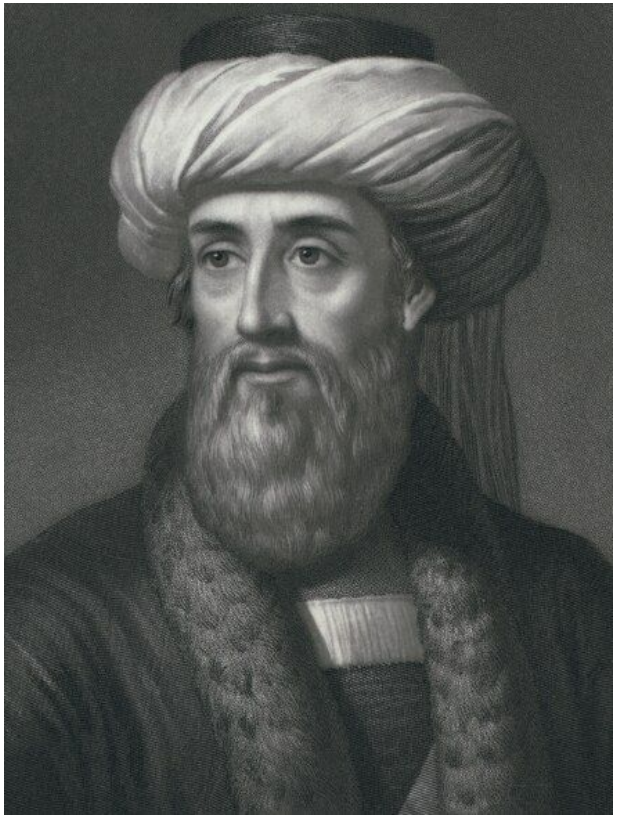

Male Lab (Red) and Male Con (Blue)
0.010
0.008
0.006
0.004
0.002
0.000
1945-1950
1950-1951
1951-1955
1955-1959
1959-1964
1964-1966
1966-1970
1970-1974
1974-1979
1979-1983
1983-1987
1987-1992
1992-1997
1997-2001
2001-2005
2005-2010
2010-2014


Fem. Lab (Red) and Fem. Con (Blue)
0.08
0.06
0.04
0.02
0.00
1945-1950
1950-1951
1951-1955
1955-1959
1959-1964
1964-1966
1966-1970
1970-1974
1974-1979
1979-1983
1983-1987
1987-1992
1992-1997
1997-2001
2001-2005
2005-2010
2010-2014

Case Study

Female Lemmas	% Parl. used more	Male Lemmas	% Parl. used more
child	1.00	argument	0.82
woman	1.00	force	0.82
health	0.94	proposition	0.76
age	0.94	corporation	0.76
mother	0.94	defence	0.71
care	0.94	army	0.71
family	0.94	doubt	0.65
husband	0.88	nuclear	0.65
elderly	0.88	British	0.65
work	0.88	parliament	0.59
help	0.88	states	0.59
parent	0.88	Europe	0.59
young	0.88	military	0.59
person	0.88	affair	0.59
girl	0.88	kingdom	0.59
baby	0.82	united	0.59
women	0.82	sense	0.53
lady	0.82	balance	0.53
home	0.82	party	0.53
maternity	0.76	industry	0.53

Introduction to Voyant Tools… and our pal Josephus

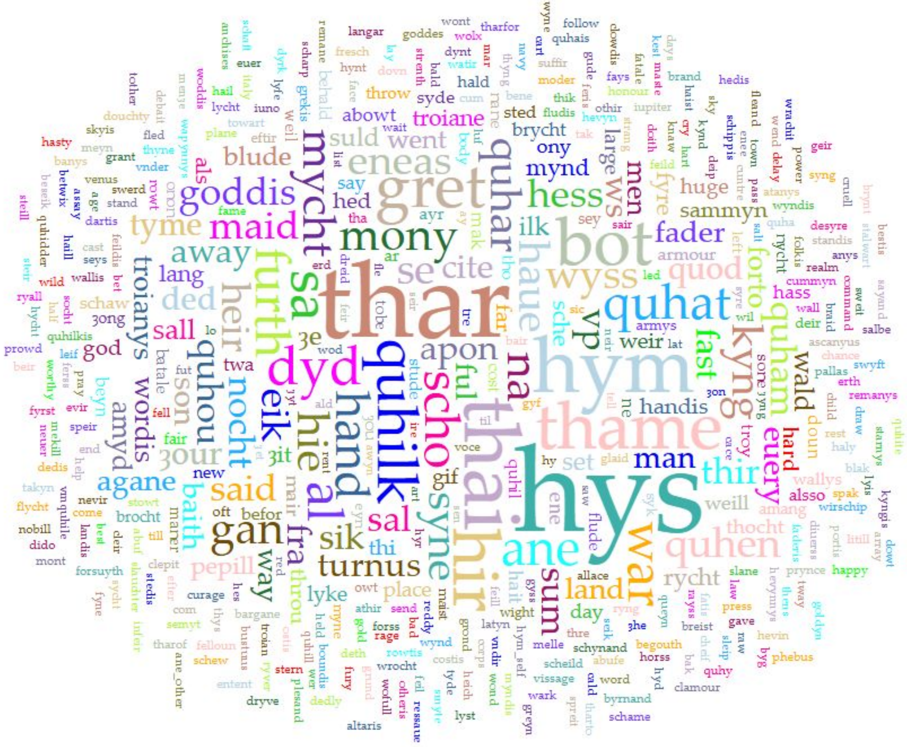


https://voyant-tools.org/

Introduction to Exercise

Try out Voyant for yourself!

1. Upload a text (or texts) to Voyant and see what you can find.
    a. Diorisis Ancient Greek Corpus
    b. Oxford Text Archive
    c. Your own text! (plain text, XML, HTML, PDF, Word, etc.)
2. Try the many tools and visualisations that Voyant offers.
    a. Reader, Contexts
    b. Collocates, Links, TermsBerry, Correlations
    c. Terms, Summary, Documents, Phrases
    d. Trends, Wordcloud, Textual Arc, Dream Scape, Veliza
3. Reflect.
    a. What worked well, what didn't?
    b. What can you now say about these texts that you couldn't before?
    c. What kinds of questions or texts do you think Voyant cannot evaluate?



Oxford Text Archive and LLDS

- National repository for literary and linguistic data.
- Oldest continuously operating archive of digital resources for literary and linguistic research.
- Over 4,000 texts from a range of time periods and a variety of languages.
- Collections are generally available and free to use.
- Committed to FAIR data principles: Findable, Accessible, Interoperable, Reusable!
- Connected to European networks.
- Check out our website at https://llds.ling-phil.ox.ac.uk/. Use the **Core Collection** - not the **Legacy Collection**.
- Have any questions? Interested in depositing? Contact me at megan.bushnell@ling-phil.ox.ac.uk.

Works Cited - Not as new as we think?

- Al-Kadit, I.A. 1992. ‘Origins of Cryptology: The Arab Contributions’, *Cryptologia*, 16.2, 97-126.
- Bjork, R.E (ed.). 2010. *The Oxford Dictionary of the Middle Ages* (Oxford: OUP).
- Bontrager, T. 1991. ‘The Development of Word Frequency Lists Prior to the 1944 Thorndike-Lorge List’, *Reading Pyschology*, 12.2, 91-116.
- Fenlon, J.F. 1913. ‘Concordances of the Bible’, in *Catholic Encyclopedia: An International Work of Reference on the Constitution, Doctrine, Discipline, and History of the Catholic Church*, ed. C.G. Herbermann et al., 15 vols (New York: The Encyclopedia Press), iv.
- Fries, C.C. 1940. *English Word Lists: A Study of Their Adaptability for Instruction* (District of Columbia: American Council on Education).
- Greetham, D. 2013. ‘A History of Textual Scholarship’, in *Cambridge Companion to Textual Scholarship*, ed. by N. Freistat and J. Flanders (Cambridge: CUP), pp. 16-41.
- ‘History of the OED’. n.d. *Oxford English Dictionary* (Oxford: OUP).
- McEnery, T. and A. Hardie. 2013. ‘The History of Corpus Linguistics’, in *The Oxford Handbook of the History of Linguistics*, ed. by K. Allan (Oxford: OUP).
- Rouse, R.H and M.A. Rouse. 1974. ‘Biblical Distinctions in the Thirteenth Century’, *Archives d’histoire doctrinale et littéraire du Moyen Age*, 41, 27-37.

Works Cited - Case Study 2

Early printed books

- Ascensius, Jodocus Badius. 1501. [Commentary], in *Opera cum variorum commentariis*, ed. by Jodocus Badius Ascensius (Paris: Ascensius & Jean Petit). UB Freiburg, Freiburg, Ink 4. D 7672, in *Freiburger historiche Bestände—digital*.
- Servius. 1501. [Commentary], in *Opera cum variorum commentariis*, ed. by Jodocus Badius Ascensius (Paris: Ascensius & Jean Petit). UB Freiburg, Freiburg, Ink 4. D 7672, in *Freiburger historiche Bestände—digital*.
- Virgil. 1501. *Opera cum variorum commentariis*, ed. by Jodocus Badius Ascensius (Paris: Ascensius & Jean Petit). UB Freiburg, Freiburg, Ink 4. D 7672, in *Freiburger historiche Bestände—digital*.

Primary sources

- Dryden, John. [n.d.]. *Aeneid*, in *Perseus Digital Library*.
- Williams, T. 1910. *The Æneid of Virgil: Translated into English Verse* (Boston: Houghton Mifflin), in *Perseus Digital Library.*

Secondary sources

- Bawcutt, P. 1973. ‘Gavin Douglas and the Text of Virgil’, *Edinburgh Bibliographical Transactions*, 4.6, 211-231.
- Bushnell, M. 2021. ‘Equivalency, Page Design, and Corpus Linguistics’ (unpublished doctoral thesis, University of Oxford).
- Bushnell M. 2025. ‘Creating a Literary Koine: How Gavin Douglas Translates Repetition in the Eneados’, in *Linguistic Fragmentation and Cultural Inclusion in the Middle Ages: Translation, Plurilingualism, Multilingualism*, ed. by D. Bertagnolli and A. Zironi, The Medieval Translator, 22 (Turnhout: Brepols).
- Moskalew, W. 1982. *Formular Language and Poetic Design in the ‘Aeneid’*, Mnemosyne Supplements, 73 (Leiden: Brill).