A gentle introduction to Digital Humanities and stylometry

Artjoms Šeļa 04.11.2019

Part I: The Big Tent

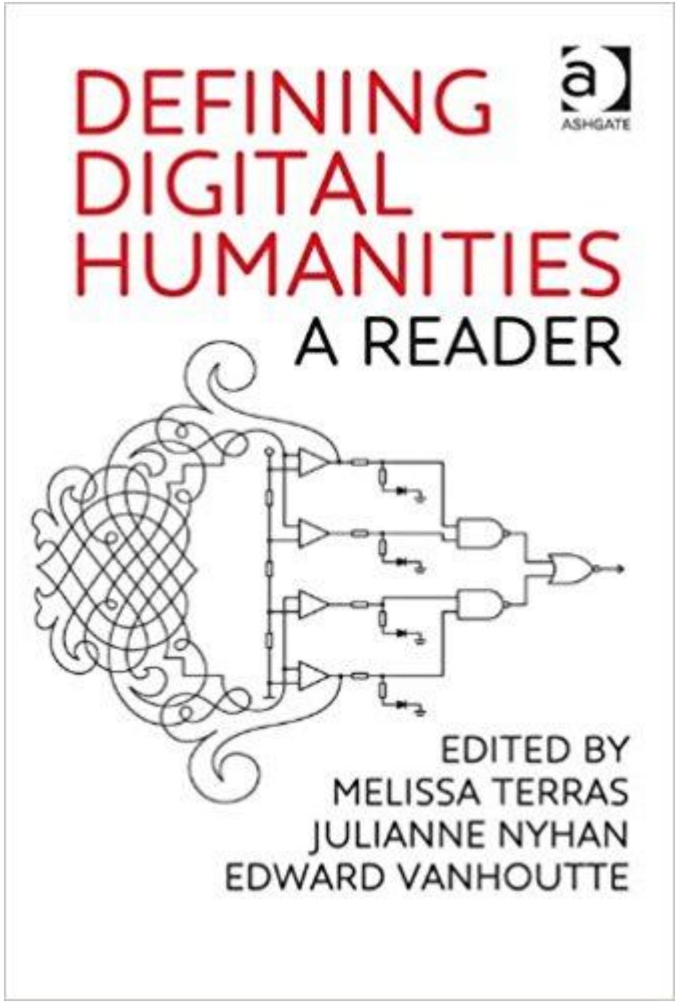
The Big Tent

- Digital Humanities (DH) : a research field(s) and research practices that emerge on the intersection of humanities, computer science and digital media.
- DH is an **umbrella-term**:
  - Quantitative/computational studies of culture;
  - Spatial humanities (Geographic Information Systems, GIS);
  - Digital critical editions;
  - Digital preservation and forensics;
  - Heritage institutions: museum and archive data curation;
  - Data journalism / data essays (e.g. *Pudding*);
  - New media studies (incl. Game studies);
  - “Digital” literature and art (digital-born objects);
  - Media-archeology;
  - ...

The many notions of DH

“A term of tactical convenience”
(Matthew Kirschenbaum)

http://whatisdigitalhumanities.com/



DEFINING DIGITAL HUMANITIES
A READER

EDITED BY
MELISSA TERRAS
JULIANNE NYHAN
EDWARD VANHOUTTE

Digital Humanities Literacy Guidebook (2019)

https://cmu-lib.github.io/dhlg/topics/

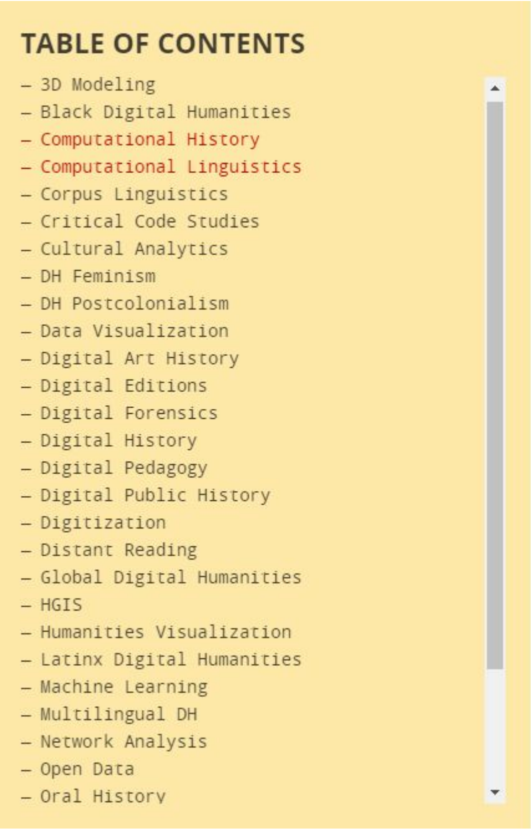
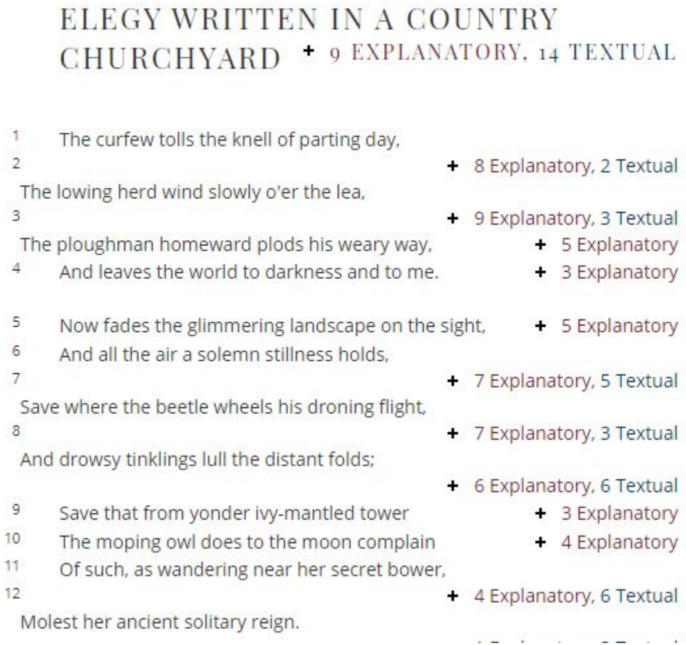


TABLE OF CONTENTS

– 3D Modeling
– Black Digital Humanities
– Computational History
– Computational Linguistics
– Corpus Linguistics
– Critical Code Studies
– Cultural Analytics
– DH Feminism
– DH Postcolonialism
– Data Visualization
– Digital Art History
– Digital Editions
– Digital Forensics
– Digital History
– Digital Pedagogy
– Digital Public History
– Digitization
– Distant Reading
– Global Digital Humanities
– HGIS
– Humanities Visualization
– Latinx Digital Humanities
– Machine Learning
– Multilingual DH
– Network Analysis
– Open Data
– Oral History

- **Expansion of existing fields of research:** *Black Digital Humanities; DH Postcolonialism, DH feminism, Digital Art History, Corpus Linguistics*
- **Broad overarching concepts:** *Distant Reading Cultural Analytics, Open Data*
- **Text processing:** *digitization, text encoding initiative (TEI)*
- **Overarching methodology and mega-frameworks:** *Geo-Information Systems, Data Visualization, Machine Learning, Text Mining, 3D modeling;*
- **Specific fields that emerged to deal with a digital-born world:** *web-archives, critical code reading, digital forensics*

Scholarly Editions



ELEGY WRITTEN IN A COUNTRY CHURCHYARD + 9 EXPLANATORY, 14 TEXTUAL

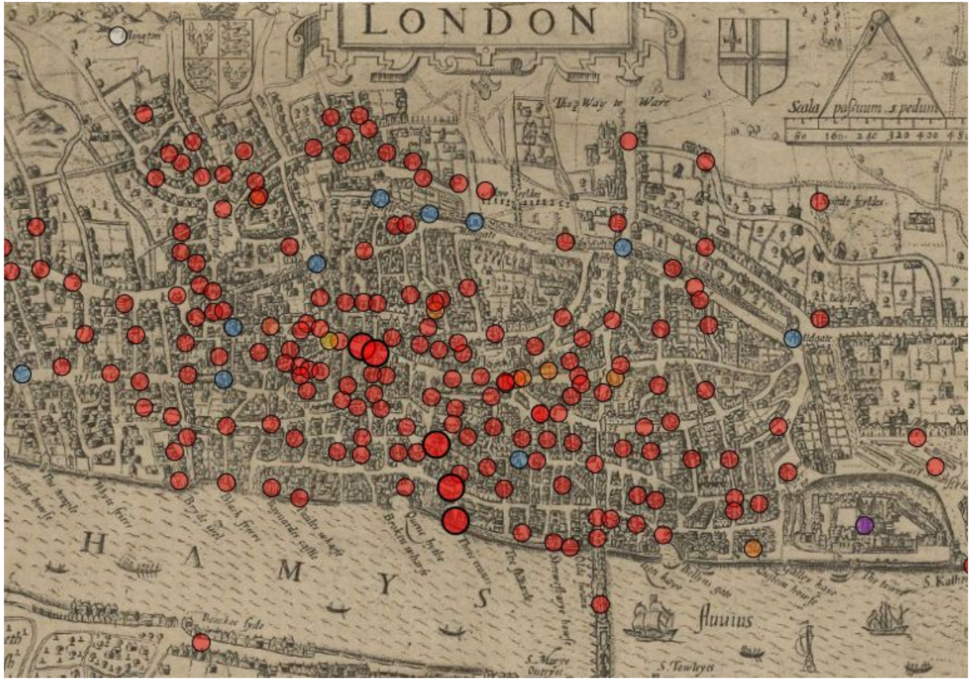
The curfew tolls the knell of parting day,
+ 8 Explanatory, 2 Textual
The lowing herd wind slowly o'er the lea,
+ 9 Explanatory, 3 Textual
The ploughman homeward plods his weary way, + 5 Explanatory
And leaves the world to darkness and to me. + 3 Explanatory

Now fades the glimmering landscape on the sight, + 5 Explanatory
And all the air a solemn stillness holds,
+ 7 Explanatory, 5 Textual
Save where the beetle wheels his droning flight,
+ 7 Explanatory, 3 Textual
And drowsy tinklings lull the distant folds;
+ 6 Explanatory, 6 Textual
Save that from yonder ivy-mantled tower + 3 Explanatory
The moping owl does to the moon complain + 4 Explanatory
Of such, as wandering near her secret bower,
+ 4 Explanatory, 6 Textual
Molest her ancient solitary reign.

- Digital medium serves perfectly in reshaping critical apparatus technology, representing different modes of reading and variants collation;
- Utilizes digitization projects, text encoding, html, techniques for text collation

https://www.thomasgray.org

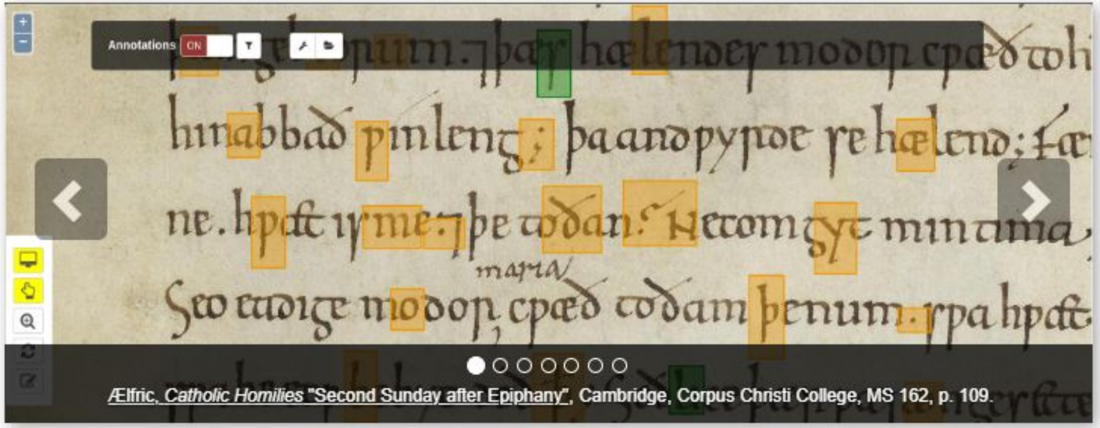
Increase of scale: “edition” of the “Grub street”



- Edition transformed to a research platform, texts gathered and linked to represent a literary market and situation in Alexander Pope’s London

http://grubstreetproject.net/

Decrease of scale: paleography

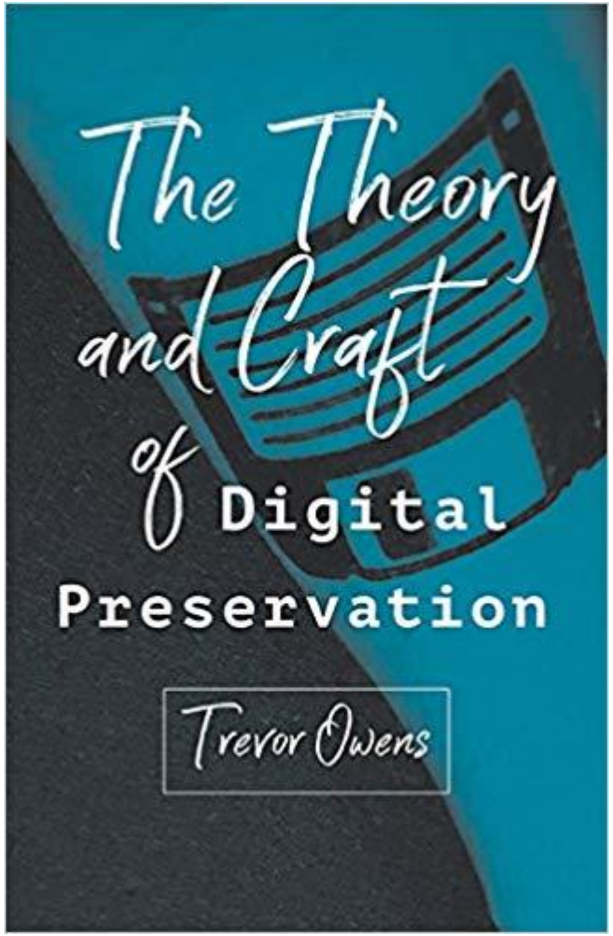


Ælfric, *Catholic Homilies* "Second Sunday after Epiphany", Cambridge, Corpus Christi College, MS 162, p. 109.

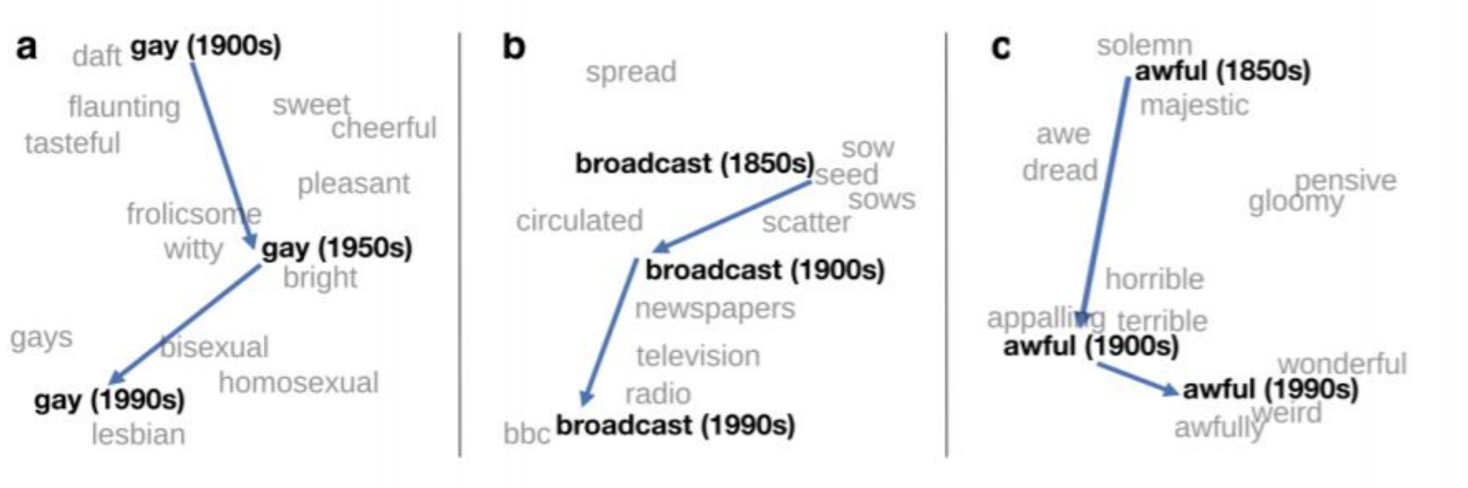
- digipal.eu
- micro-level encoding of manuscripts & characters & their features

Preservation and archives

- GLAM (Galleries, Libraries, Archives, Museums) engage in massive digitization projects
- But how to preserve a “digital-born” information?
- **Digital preservation**: a field that draws on media-archeology & digital forensics to understand the nature of digital data and arrive at long-term solutions for it’s curation, preservation and establishing access;
- Versions of videogames, old hardware, digital manuscripts, personal digital archives, databases…

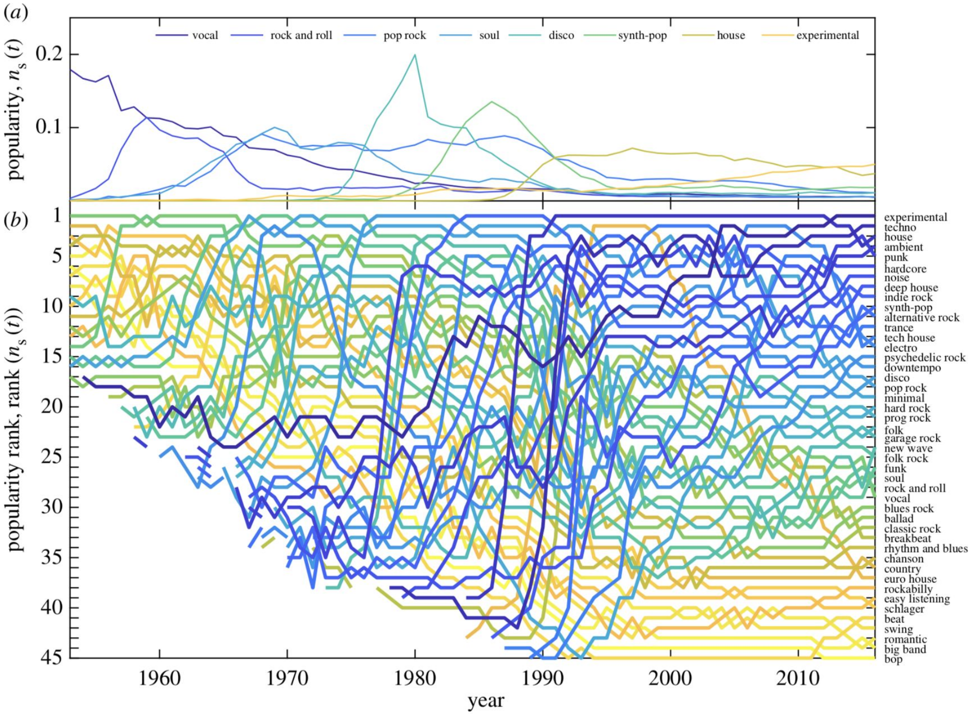


Corpus and historical linguistics



William L. Hamilton, Jure Leskovec, Dan Jurafsky. (2016) *Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change*

New modes of studying cultural change



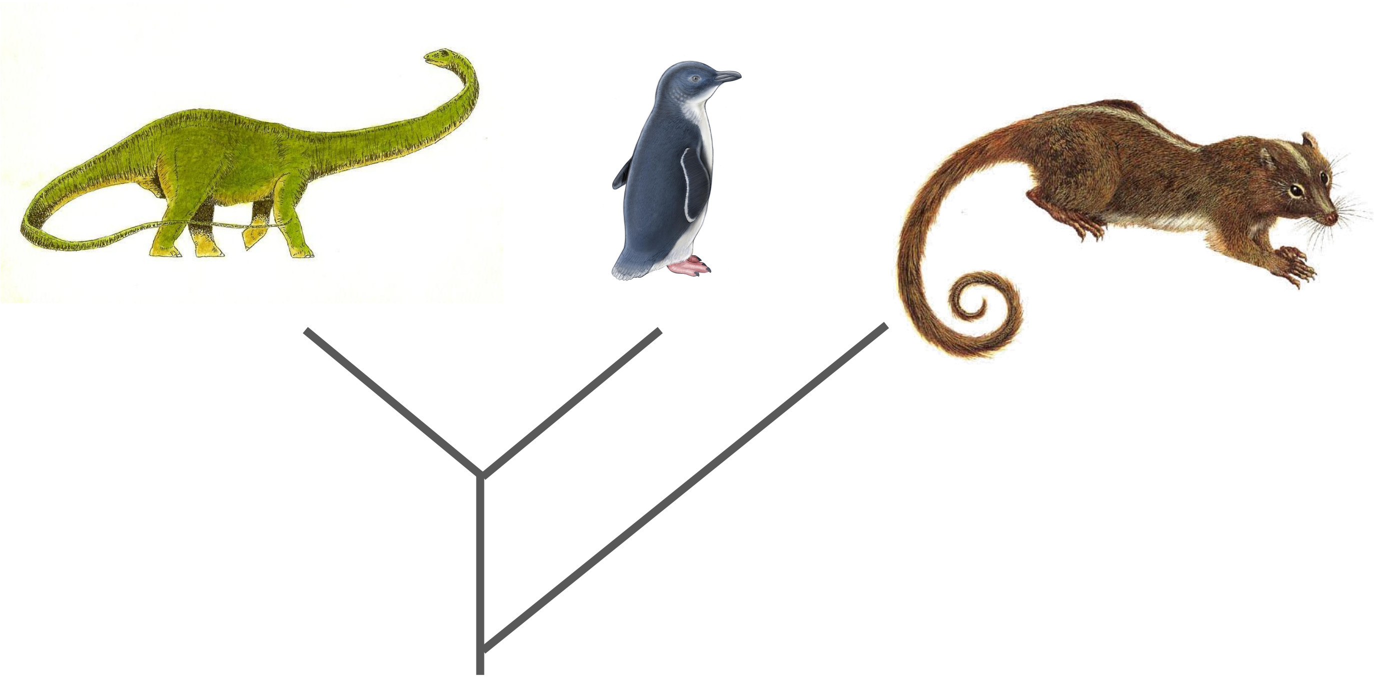
- Return of the empiricism in the humanities;
- Quantitative framework re-established our understanding of our objects, pushes us to model it, extract relevant features, test and retest global assumptions

Part II: stylometry with R

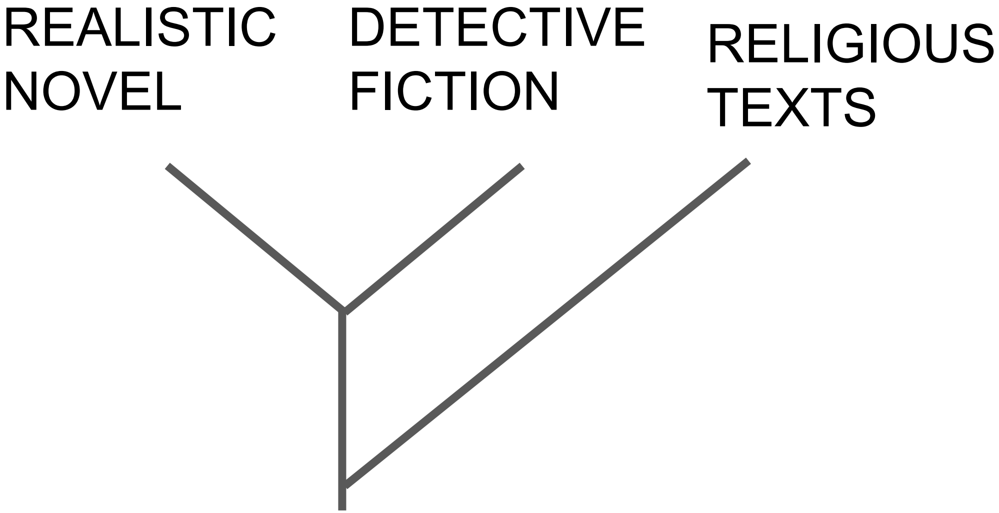
How similar? How different?

Think in terms of distances: what evolutionary path each species has undergone

Think in terms of distances: what evolutionary path each species has undergone



Continue thinking in terms of distances

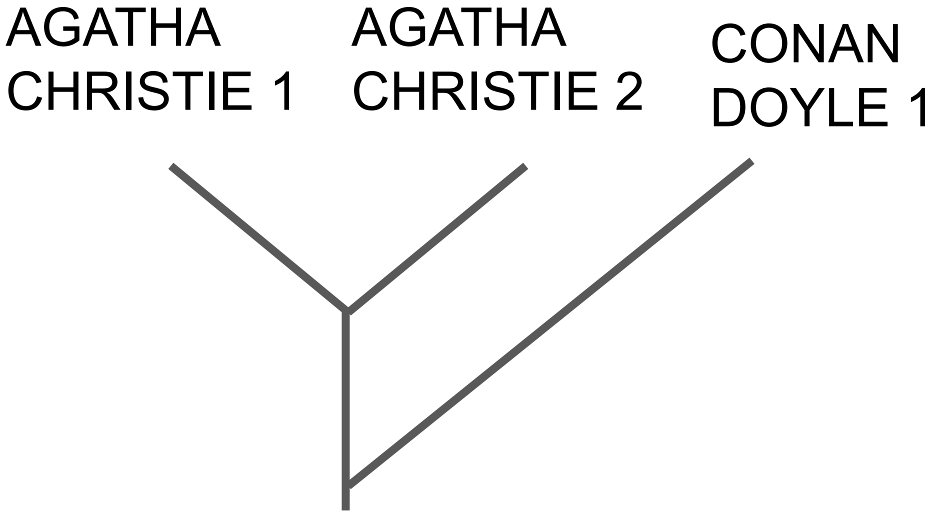


REALISTIC NOVEL

DETECTIVE FICTION

RELIGIOUS TEXTS

Continue thinking in terms of distances

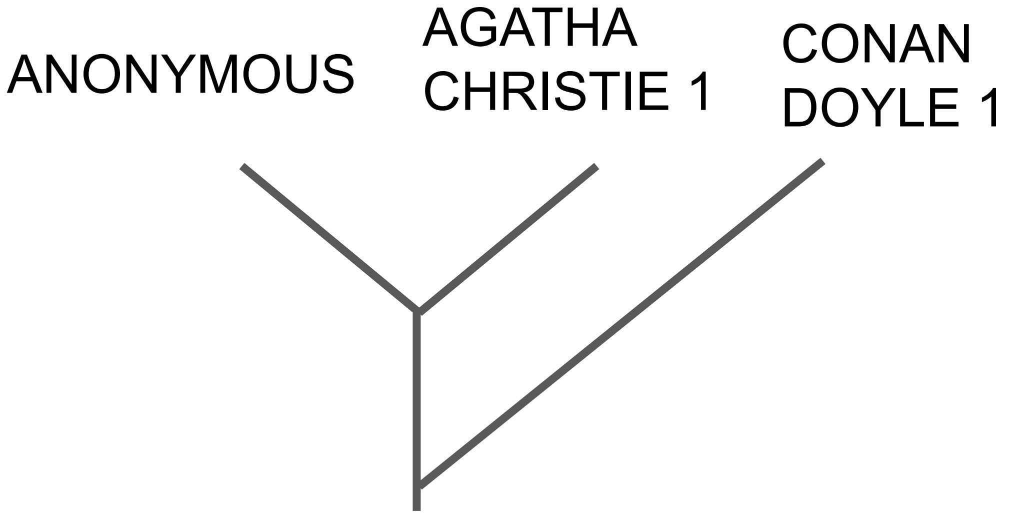


AGATHA CHRISTIE 1

AGATHA CHRISTIE 2

CONAN DOYLE 1

Continue thinking in terms of distances

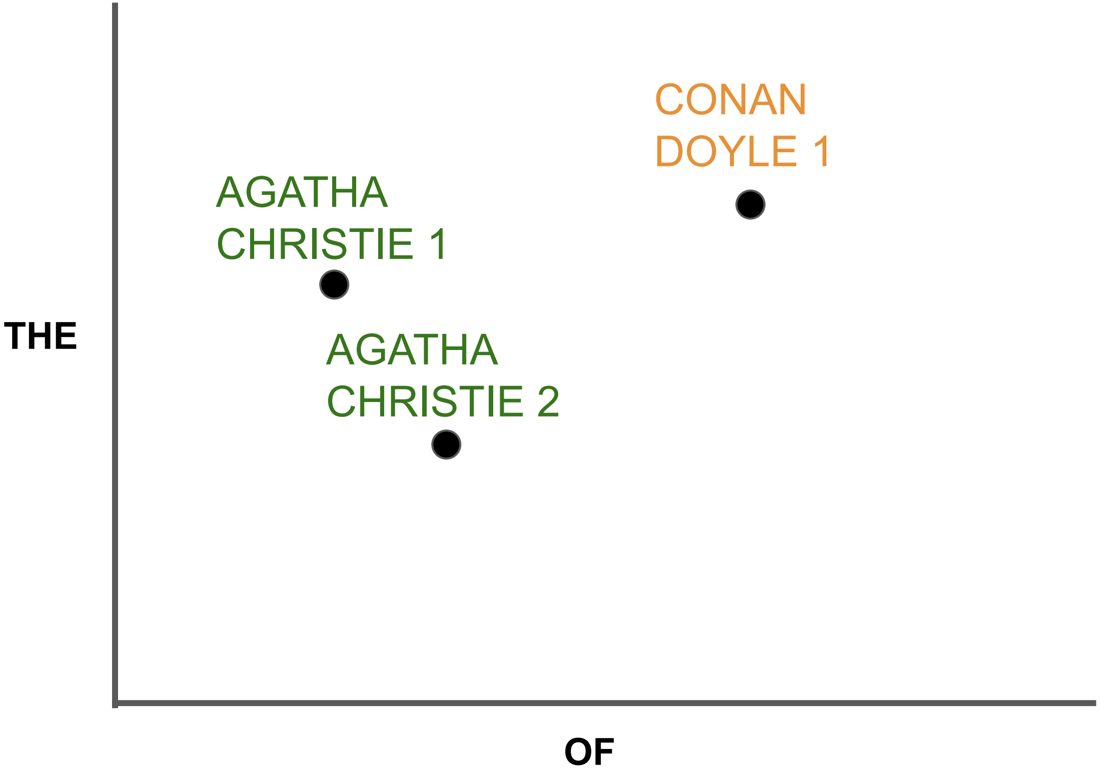


ANONYMOUS

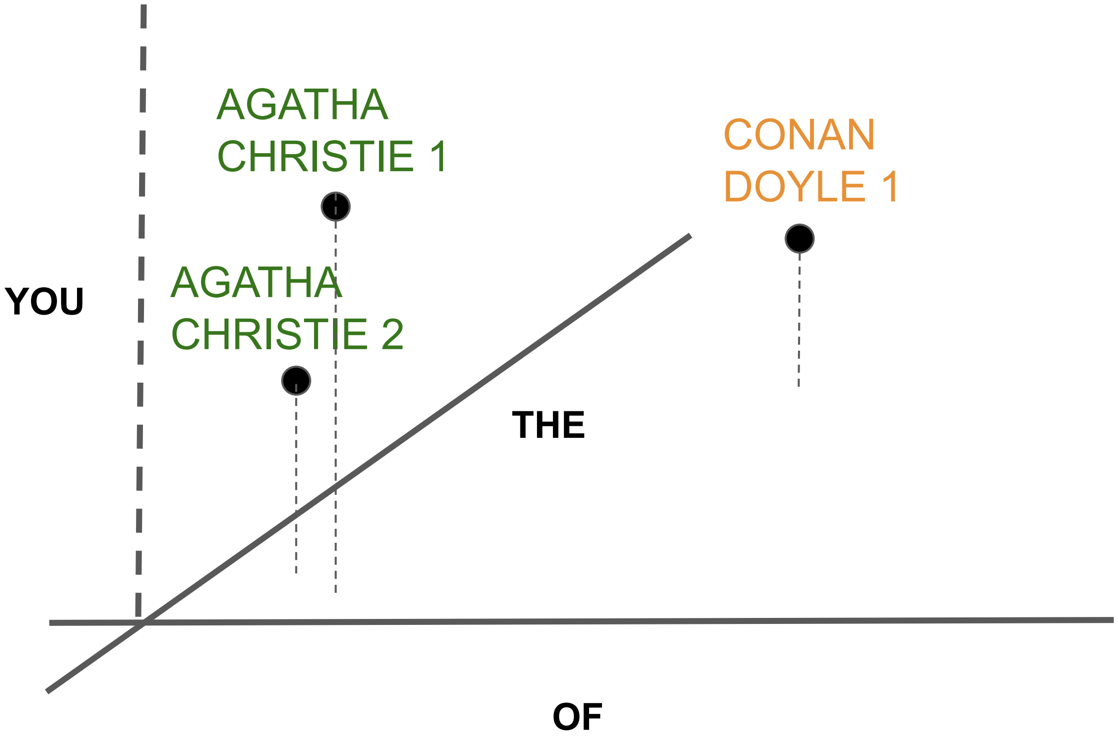
AGATHA CHRISTIE 1

CONAN DOYLE 1

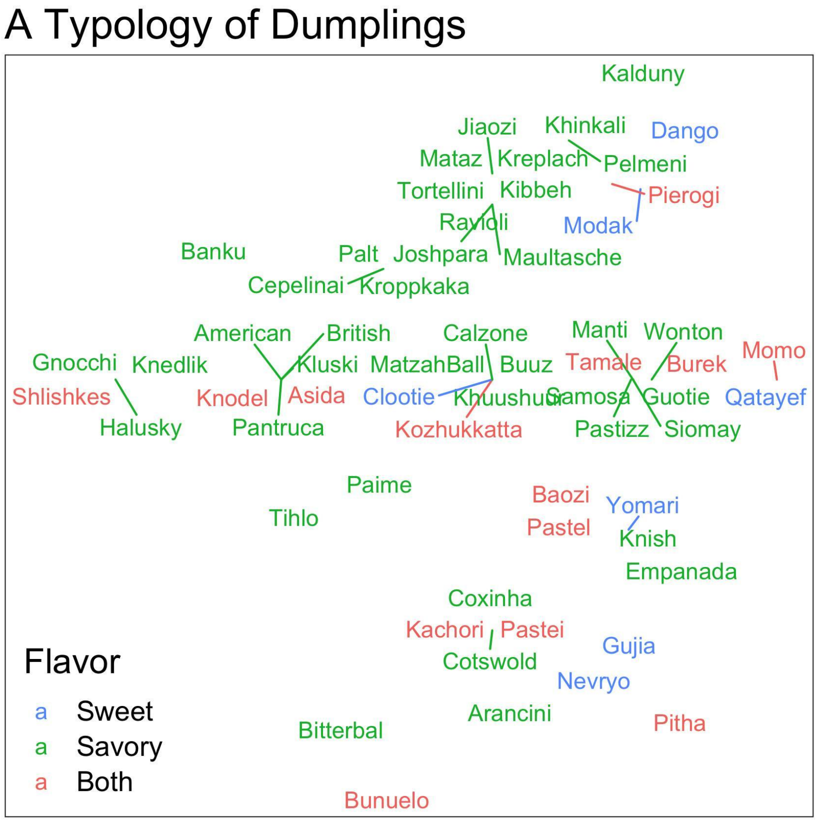
Continue thinking in terms of distances - 2D



Continue thinking in terms of distances - 3D

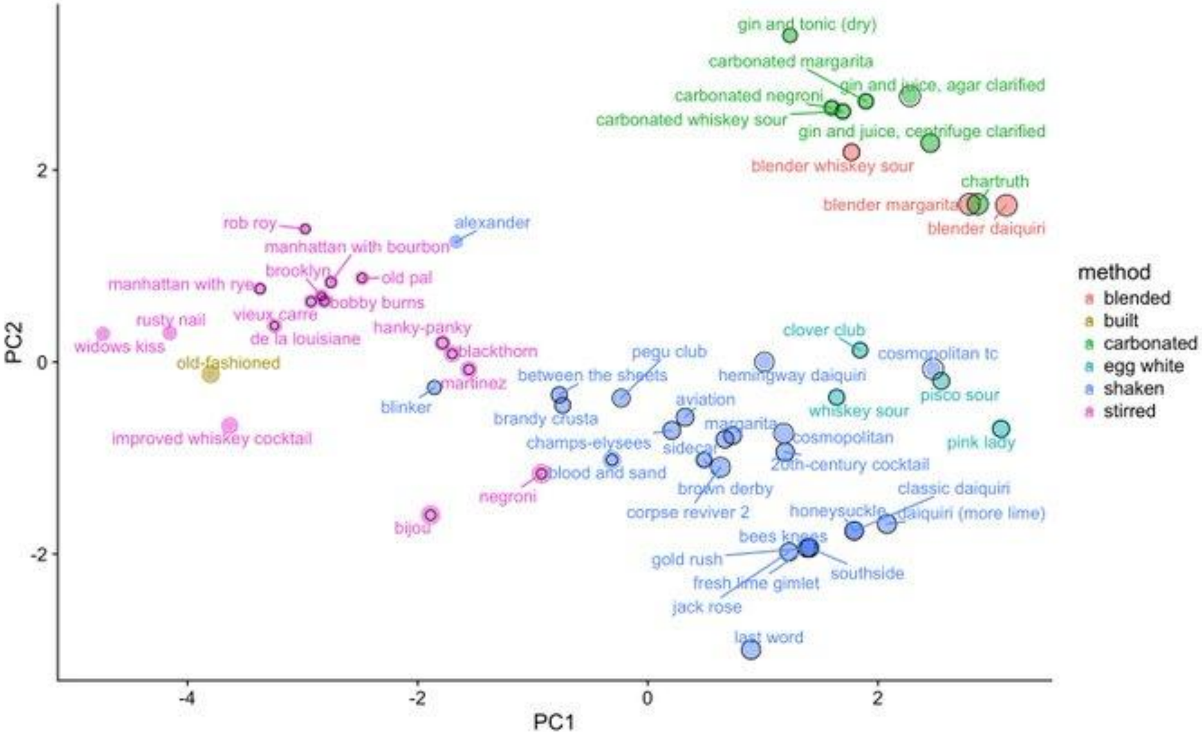


Multivariate analysis in n-dimensions



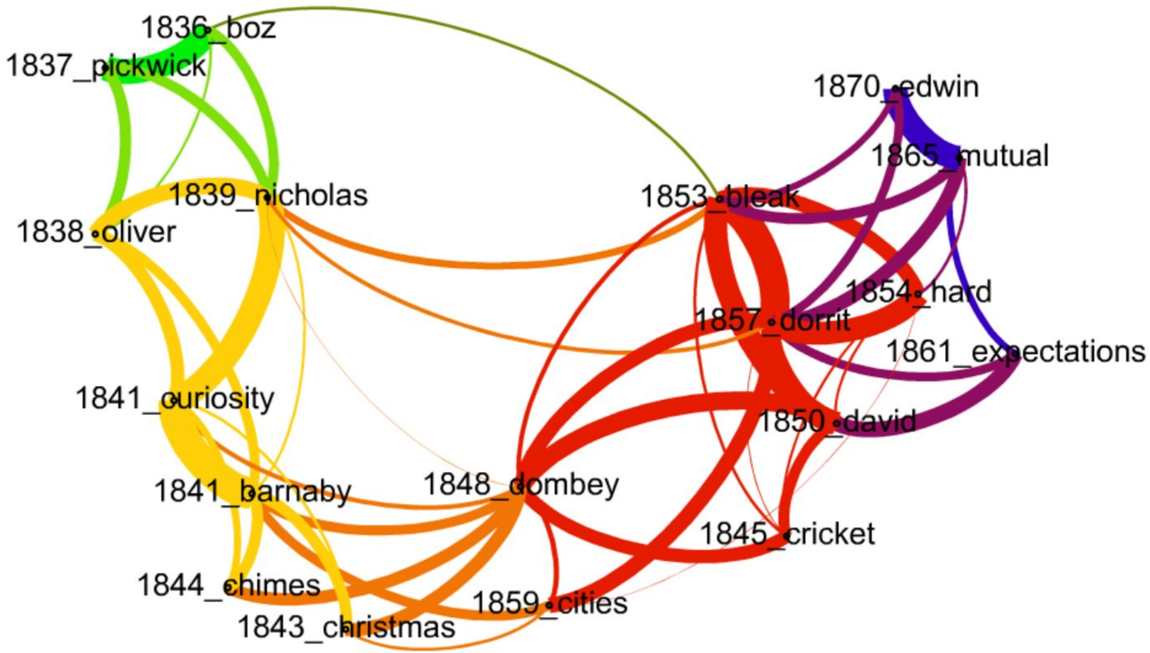
- Dumplings around the world coded as 15 features…
- Multi-Dimensional Scaling (MDS): projection on 2D
- Author: Calle Börstell (@c_borstell)

Multivariate analysis in n-dimensions



- Cocktails around the world coded as: “start/finish volume”, “acidity”, “ABV”, “sugar content”, “dilution ratio”
- Author: Harold Pimentel (@hjpimentel)

Multivariate analysis in n-dimensions



- C. Dickens novels coded in 100 most frequent words
- Distances between texts represented as edges in a network
- Author: Jan Rybicki

Stylometry tries to address the questions of style differences in a quantitative way

Stylometry tries to address the questions of style differences in a quantitative way

A “style” in this sense is some **set of quantifiable textual features that show unique behavior in different texts**

Stylometry tries to address the questions of style differences in a quantitative way

A “style” in this sense is some **set of quantifiable textual features that show unique behavior in different texts**

Many features were proposed: word lengths, sentence lengths, word frequencies, character n-grams, part-of-speech tags, versification features like rhythm and rhyme patterns…

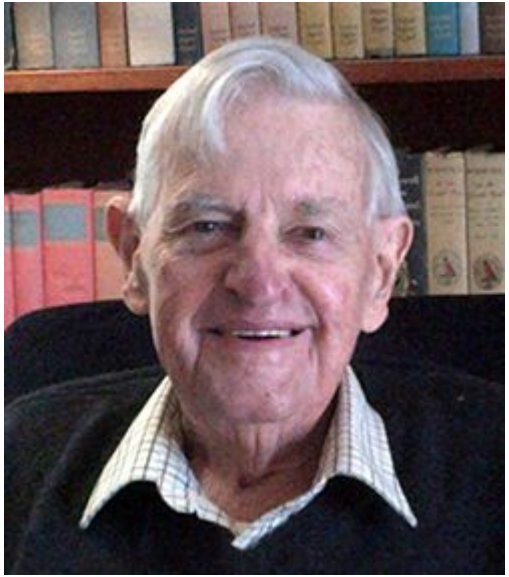
Stylometry tries to address the questions of style differences in a quantitative way

A “style” in this sense is some **set of quantifiable textual features that show unique behavior in different texts**

Many features were proposed: word lengths, sentence lengths, word frequencies, character n-grams, part-of-speech tags, versification features like rhythm and rhyme patterns…

Authorship attribution is a sub-problem of stylometry: **how unique patterns of textual features help to catch author identity**

Distances between texts: Burrow’s delta (2002)



John Burrows

$$\Delta = \sum_{i=1}^{n} \frac{|z(x_i) - z(y_i)|}{n}$$

On the distances

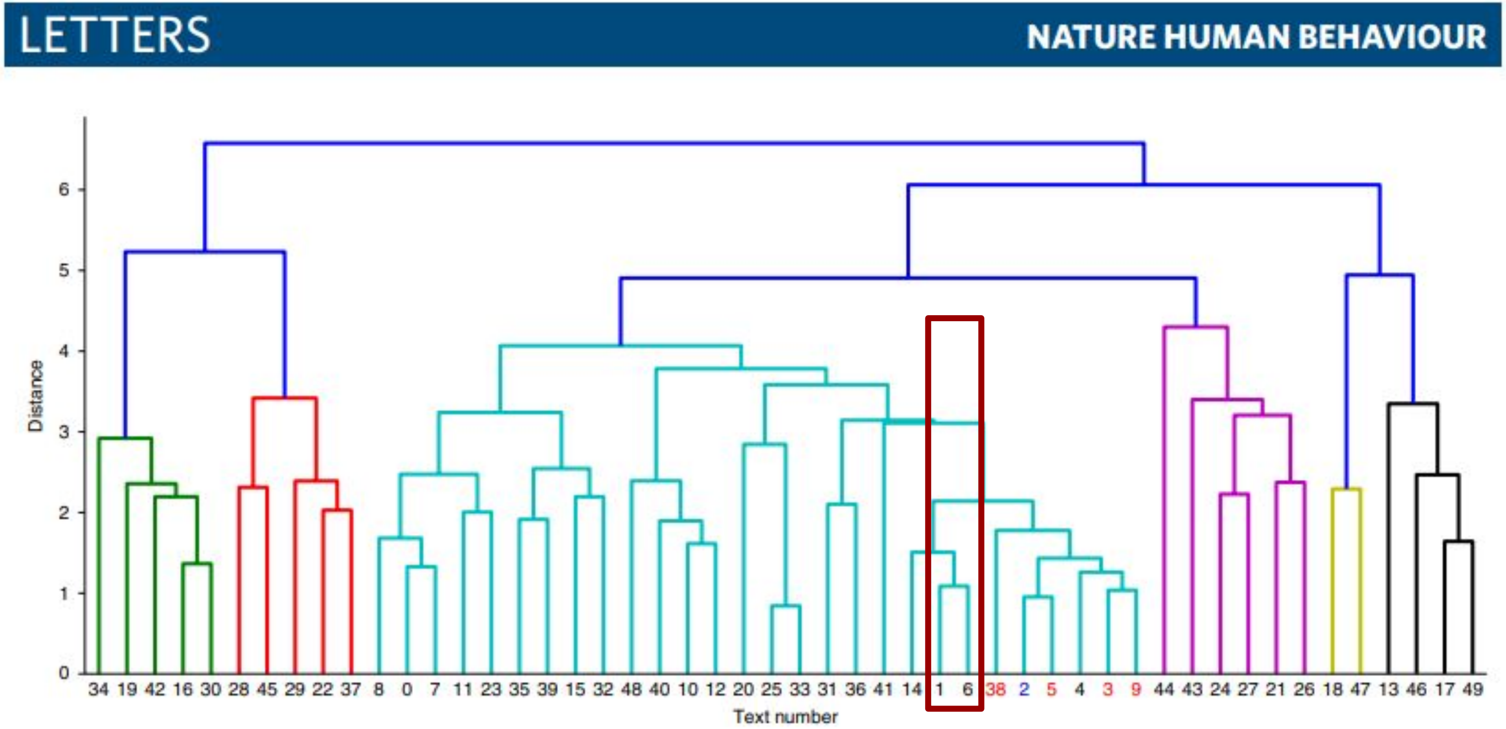
http://versologie.cz/talks/2017chicago/

Petr Plecháč (Czech Academy of Sciences)

“stylo” package for R

Eder, M., Rybicki, J. and Kestemont, M. (2016).
Stylometry with R: a package for computational text analysis. R Journal 8(1): 107-121.

Authorship of Beowulf?



Neidorf, L., Krieger, M. S., Yakubek, M., Chaudhuri, P., & Dexter, J. P. (2019). *Large-scale quantitative profiling of the Old English verse tradition. Nature Human Behaviour.* doi:10.1038/s41562-019-0570-1

For web-based distance calculation you may want to try out “Lexos” tool