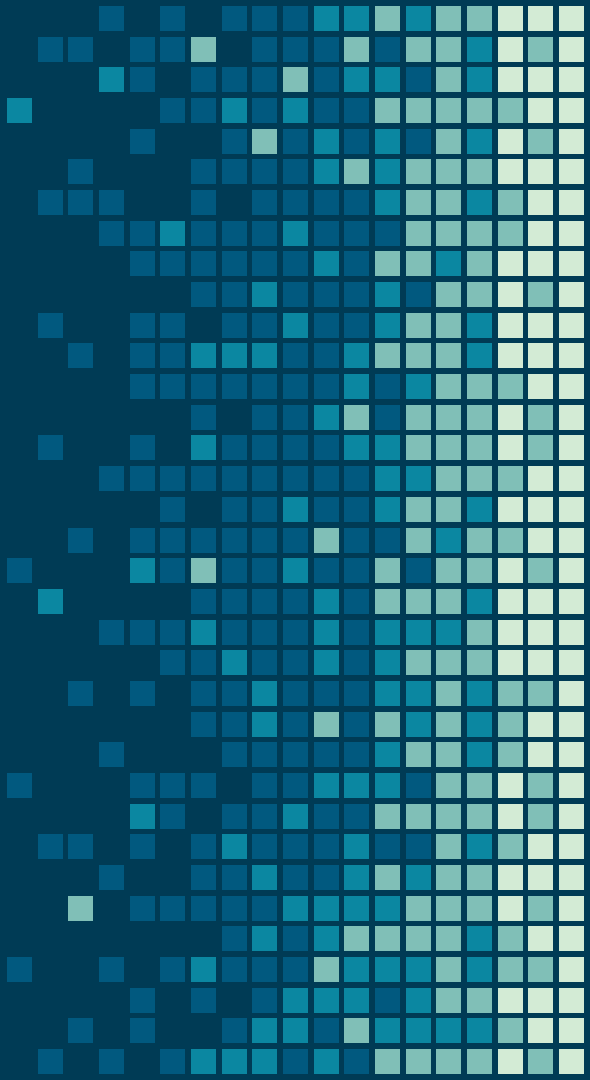
Introduction to stylometry and textual analysis

Joanna Byszuk
(Institute of Polish Language of the Polish Academy of Sciences)

Introduction to stylometry



Distant reading

- **Analysing big literary collections** based on not text but metadata / research literature etc.

(as conceptualized by Franco Moretti 2000)

- **Analysing language / literature "objectively" – at a distance**, looking at particular features within the texts

(Mendenhall 1887, Lutosławski 1890)

(but also Lorenzo Valla 1440, Augustus de Morgan 1851)

What is stylometry?

Stylometry =
use of quantitative methods
to examine similarities and differences
within a group of [texts]

Stylometry is related to

- Computational & Forensic Linguistics
- Network Analysis
- Natural Language Processing

How does it work?

How does stylometry work?

corpus of texts

+

distance measure

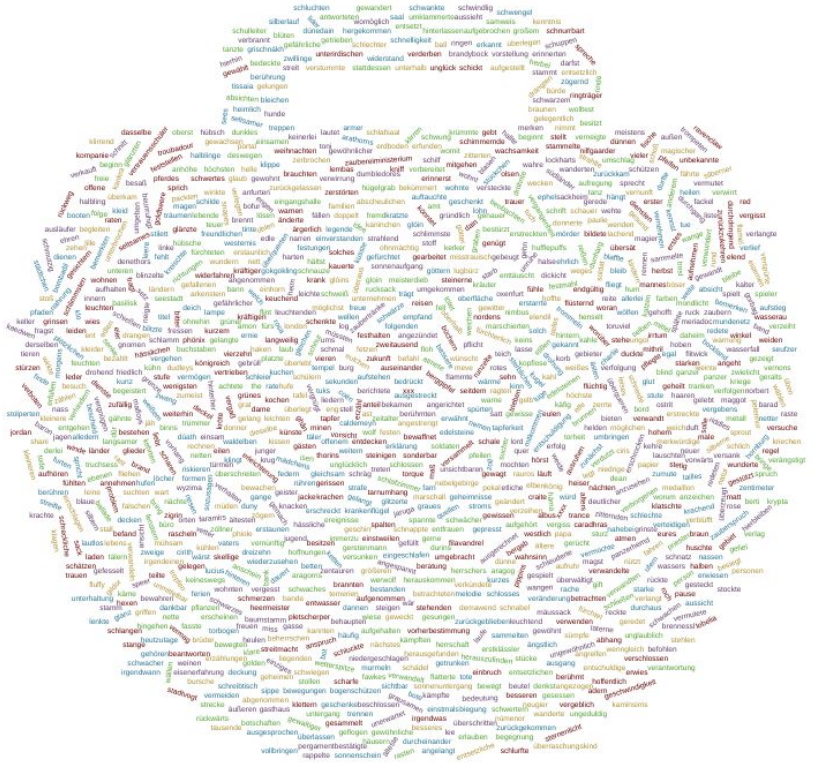
+

classification algorithm

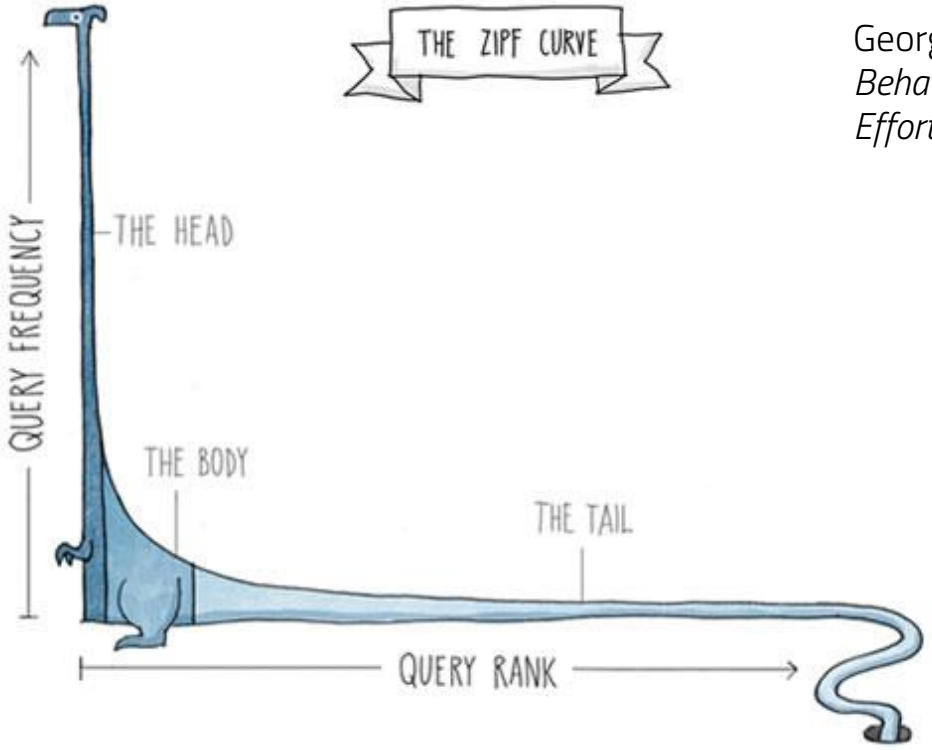
+

(visualisation)

Text as bag-of-words



Zipf's law



George K. Zipf (1949) *Human Behavior and the Principle of Least Effort*. Addison-Wesley.

Zipf's law


Am I the only one around here that tries to do things with the least effort possible and expects a good result?!
someecards
user card

What words are at the top?

	Agnes	Tenant	Emma	Pride	Sense	Jane
the	2511	5929	5204	4330	4105	7835
and	2733	6705	4878	3577	3489	6618
to	2366	5594	5186	4136	4103	5152
of	1602	3734	4292	3609	3571	4359
i	2204	6075	3191	2064	1998	7165
a	1296	2792	3126	1948	2067	4467
in	911	2021	2174	1866	1948	2762
that	776	1909	1800	1577	1383	1655
he	659	2259	1811	1338	1112	1902
was	1000	1835	2400	1847	1861	2525
it	795	2280	2529	1532	1755	2403
you	760	2844	1999	1356	1191	2971
her	750	1760	2483	2224	2543	1714

	Agnes	Tenant	Emma	Pride	Sense	Jane
the	3.67471	3.54285	3.24344	3.55705	3.43227	4.18704
and	3.99959	4.00655	3.04026	2.93847	2.91722	3.53667
to	3.46251	3.34267	3.23222	3.39768	3.43060	2.75324
of	2.34444	2.23124	2.67503	2.96476	2.98579	2.32946
i	3.22543	3.63009	1.98882	1.69556	1.67057	3.82899
a	1.89662	1.66835	1.94831	1.60026	1.72826	2.38717
in	1.33320	1.20764	1.35496	1.53290	1.62876	1.47602
that	1.13563	1.14072	1.12187	1.29549	1.15635	0.88444
he	0.96441	1.34986	1.12872	1.09915	0.92977	1.01643
was	1.46344	1.09650	1.49582	1.51729	1.55602	1.34937
it	1.16344	1.36241	1.57622	1.25852	1.46739	1.28417
you	1.11222	1.69942	1.24589	1.11394	0.99582	1.58771
her	1.09758	1.05168	1.54755	1.82699	2.12625	0.91597

For two texts *T* and *T1*, and for a set of *n* words,

$$\Delta(T,T_1) = \frac{1}{n}\sum_{i=1}^{n}\left|z(f_i(T)) - z(f_i(T_1))\right|$$

Where $z(f_x(T)) = \frac{f_x(T) - \mu_x}{\sigma_x}$;

$f_x(T)$ = raw frequency of word *x* in text T;

μ_x = mean frequency of word *x* in a collection of texts;

σ_x = standard deviation of frequency of word *x*.

= what's the difference in how two texts use a given feature, compared to its average use

E.g. …

	Agnes	Tenant	Emma	Pride	Sense	Jane
the	3.67471	3.54285	3.24344	3.55705	3.43227	4.18704
and	3.99959	4.00655	3.04026	2.93847	2.91722	3.53667
to	3.46251	3.34267	3.23222	3.39768	3.43060	2.75324
of	2.34444	2.23124	2.67503	2.96476	2.98579	2.32946
i	3.22543	3.63009	1.98882	1.69556	1.67057	3.82899
a	1.89662	1.66835	1.94831	1.60026	1.72826	2.38717
in	1.33320	1.20764	1.35496	1.53290	1.62876	1.47602
that	1.13563	1.14072	1.12187	1.29549	1.15635	0.88444
he	0.96441	1.34986	1.12872	1.09915	0.92977	1.01643
was	1.46344	1.09650	1.49582	1.51729	1.55602	1.34937
it	1.16344	1.36241	1.57622	1.25852	1.46739	1.28417
you	1.11222	1.69942	1.24589	1.11394	0.99582	1.58771
her	1.09758	1.05168	1.54755	1.82699	2.12625	0.91597

	Agnes	Pride	Jane	David	Mill	Tom	Clarissa
Tenant	0.81	1.07	0.88	0.92	0.98	1.16	1.1
Emma	1.12	0.78	1.28	1.15	1.2	1.25	1.24
Sense	1.14	0.69	1.24	1.16	1.25	1.13	1.21
Professor	1.06	1.21	0.69	0.94	1	1.27	1.3
Villette	1.07	1.26	0.65	0.91	0.96	1.28	1.3
Bleak	1.09	1.18	0.92	0.55	0.87	1.21	1.17
Hard	1.16	1.25	0.96	0.65	0.91	1.26	1.25
Wuthering	1.06	1.31	0.81	0.94	1.01	1.32	1.27
Adam	1.13	1.37	0.95	0.9	0.66	1.42	1.32
Middlemarch	1.01	1.1	0.99	0.87	0.65	1.17	1.12
Joseph	1.2	1.19	1.24	1.18	1.29	0.64	1.11
Pamela	1.15	1.24	1.27	1.19	1.26	1.11	0.67
Sentimental	1.38	1.53	1.23	1.22	1.29	1.42	1.38

What words at the top? What relevance?

- Grammatical words occupy the top of the frequency list

(Zipf, 1948)

- Grammatical words are strong predictors

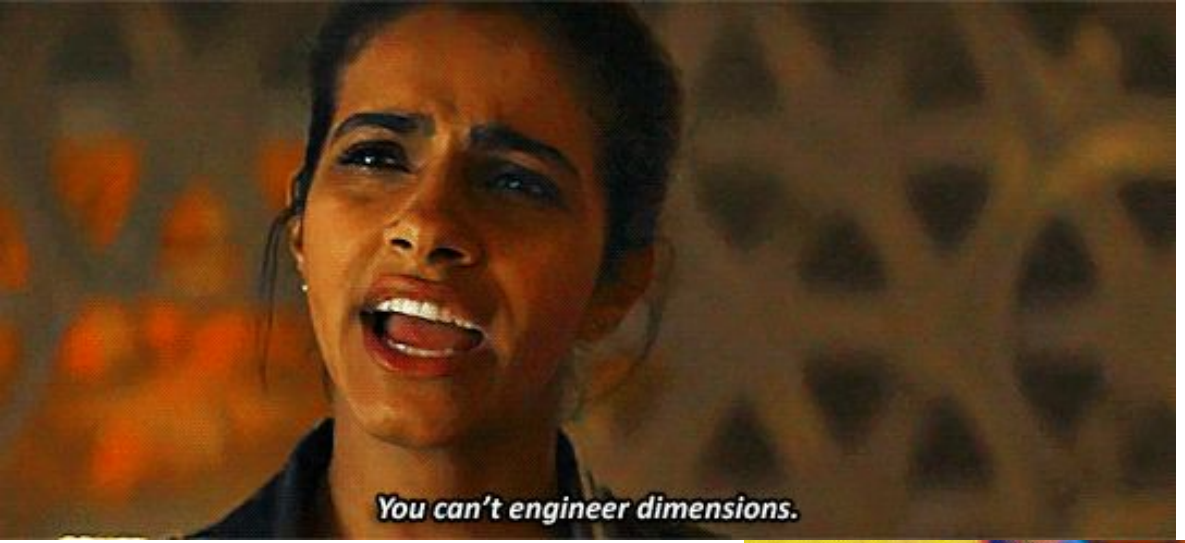
(Mosteller & Wallace, 1964)

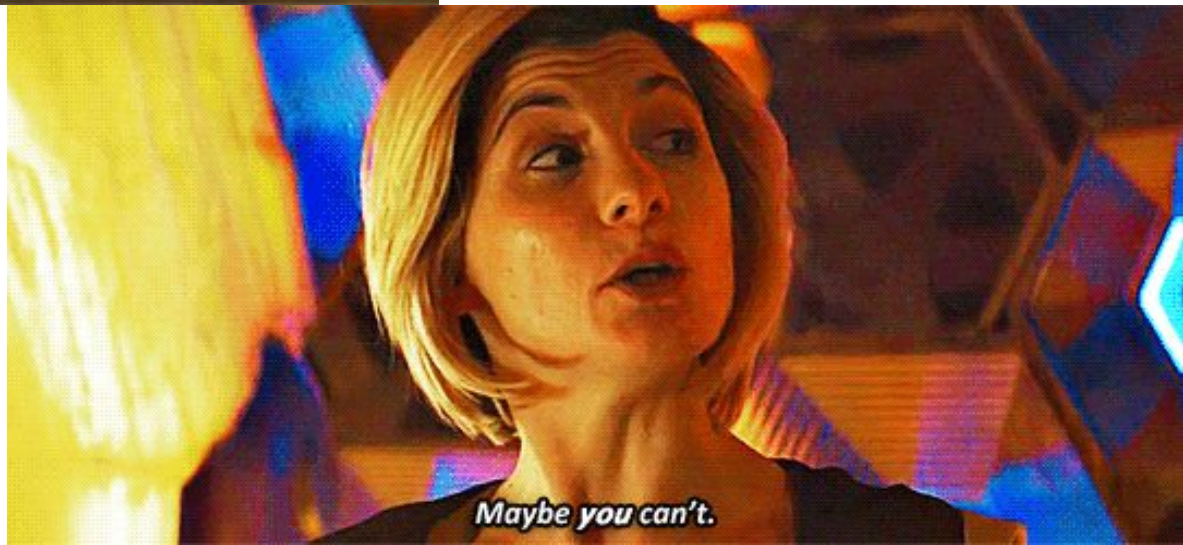
- Therefore: top N words are strong predictors

(numerous stylometrists around the world)

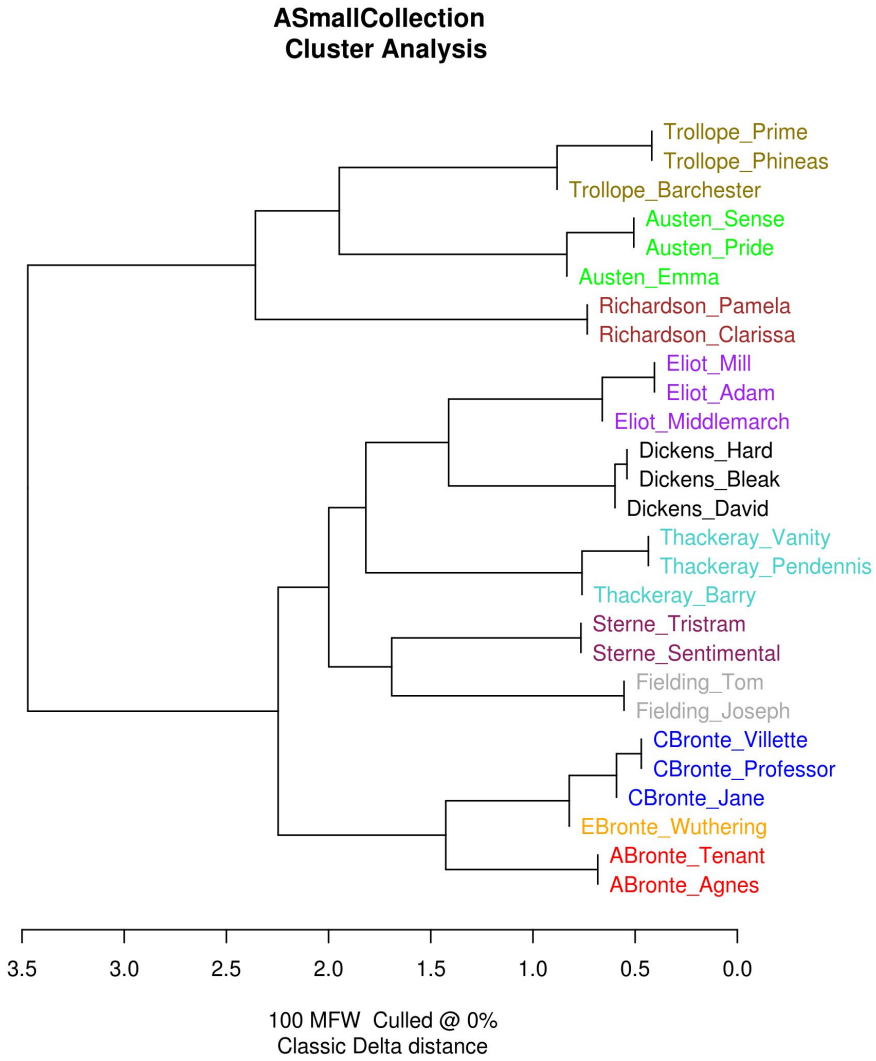
Cool, but how to actually analyse this?

Answer: dimension reduction


You can’t engineer dimensions.

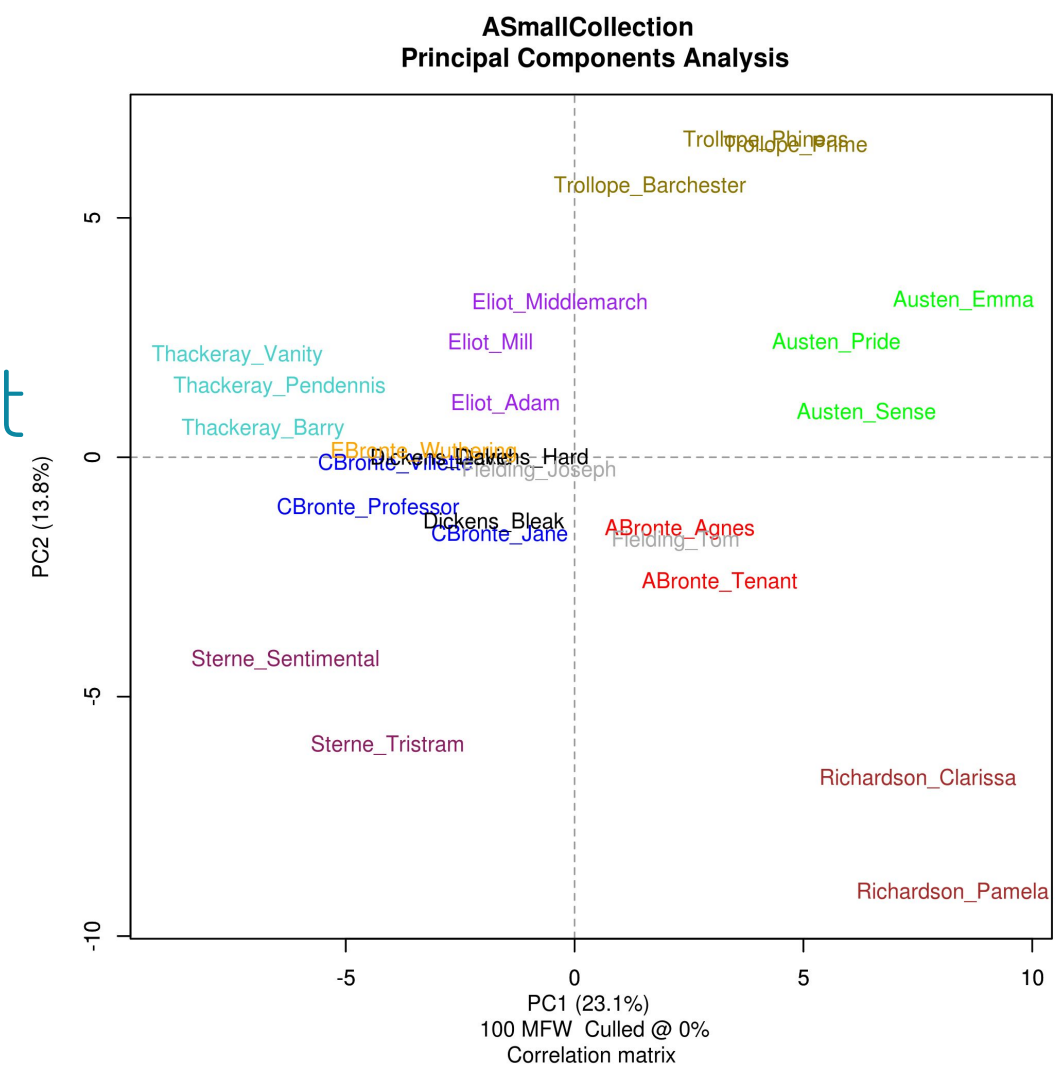

Maybe you can’t.

#1 Cluster analysis


ASmallCollection
Cluster Analysis
Trollope_Prime
Trollope_Phineas
Trollope_Barchester
Austen_Sense
Austen_Pride
Austen_Emma
Richardson_Pamela
Richardson_Clarissa
Eliot_Mill
Eliot_Adam
Eliot_Middlemarch
Dickens_Hard
Dickens_Bleak
Dickens_David
Thackeray_Vanity
Thackeray_Pendennis
Thackeray_Barry
Sterne_Tristram
Sterne_Sentimental
Fielding_Tom
Fielding_Joseph
CBronte_Villette
CBronte_Professor
CBronte_Jane
EBronte_Wuthering
ABronte_Tenant
ABronte_Agnes
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
100 MFW Culled @ 0%
Classic Delta distance

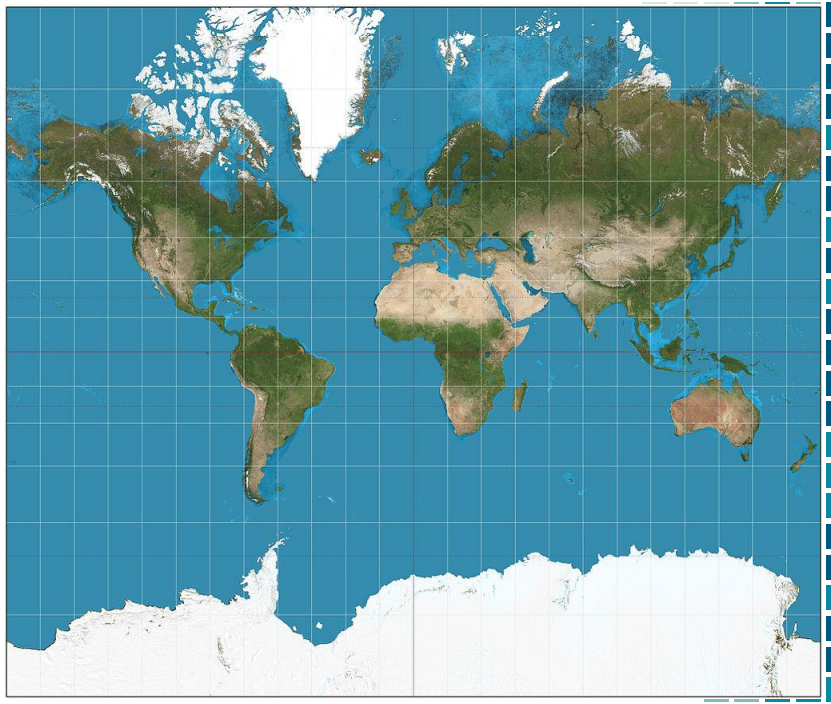
#2 PCA

PCA – Principal Component Analysis



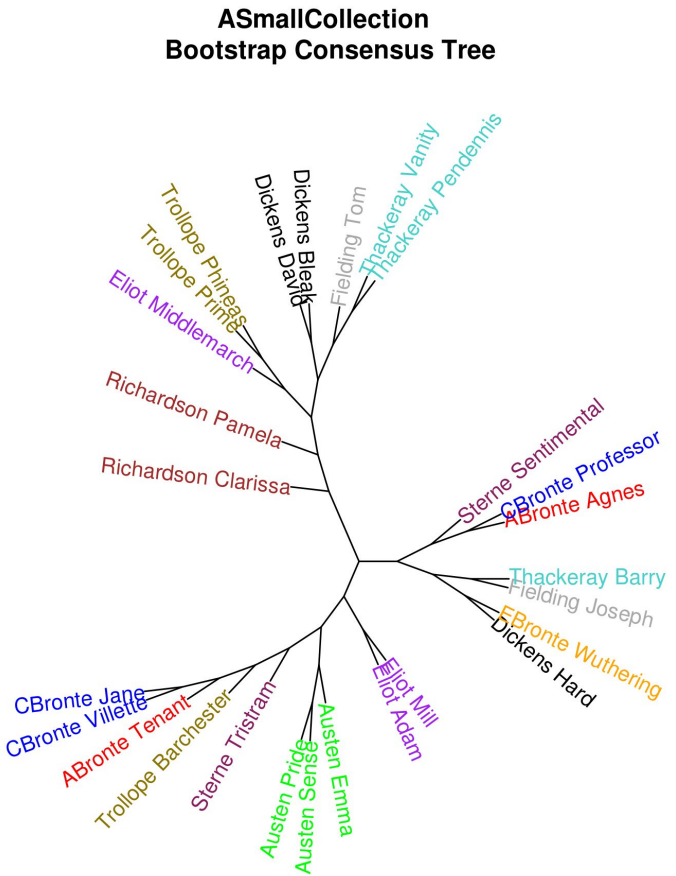
Problems with dimension reduction

Information loss

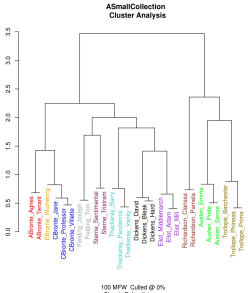


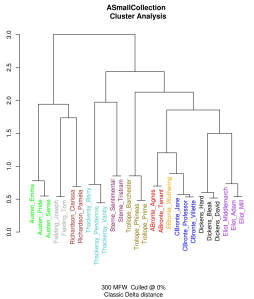
Gold standard

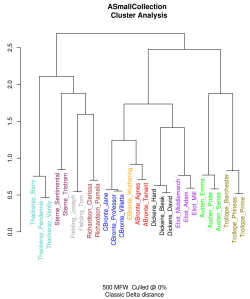
Bootstrap consensus tree

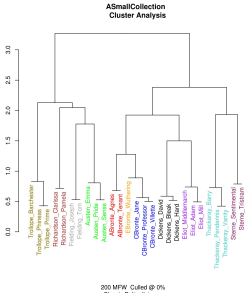


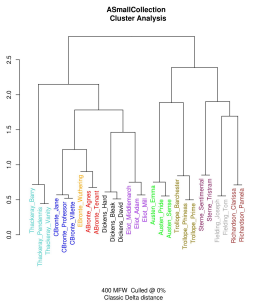
Bootstrap consensus tree – how?

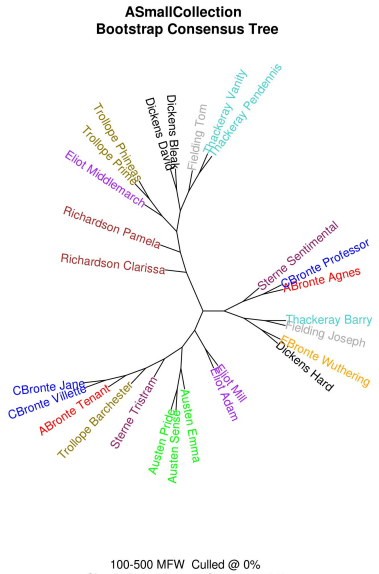




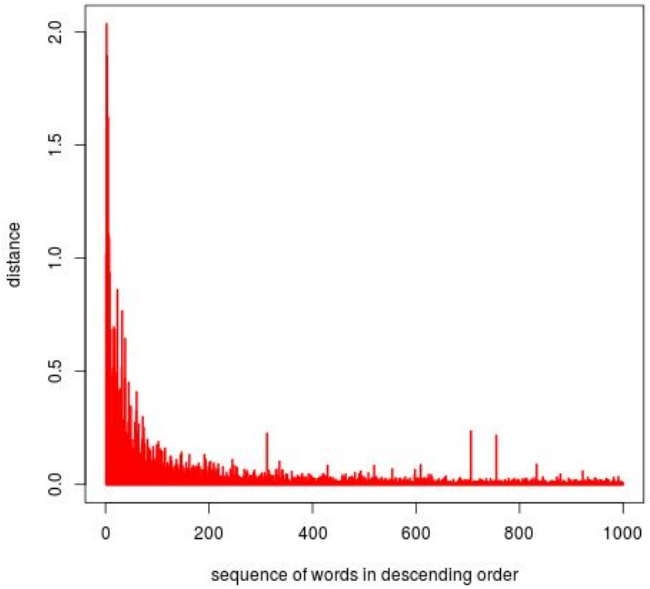






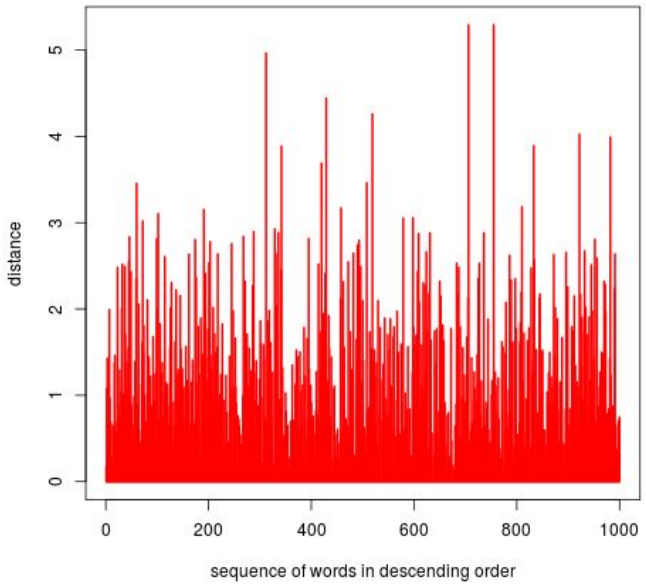


Distance measure



Classic mathematical methods, e.g. Euclidean distance?

Distance measure



Classic Burrows distance!

For two texts *T* and *T1,* and for a set of *n* words,

$$\Delta(T, T_1) = \frac{1}{n}\sum_{i=1}^{n}\left|z(f_i(T)) - z(f_i(T_1))\right|$$

Where $z(f_x(T)) = \frac{f_x(T) - \mu_x}{\sigma_x}$;

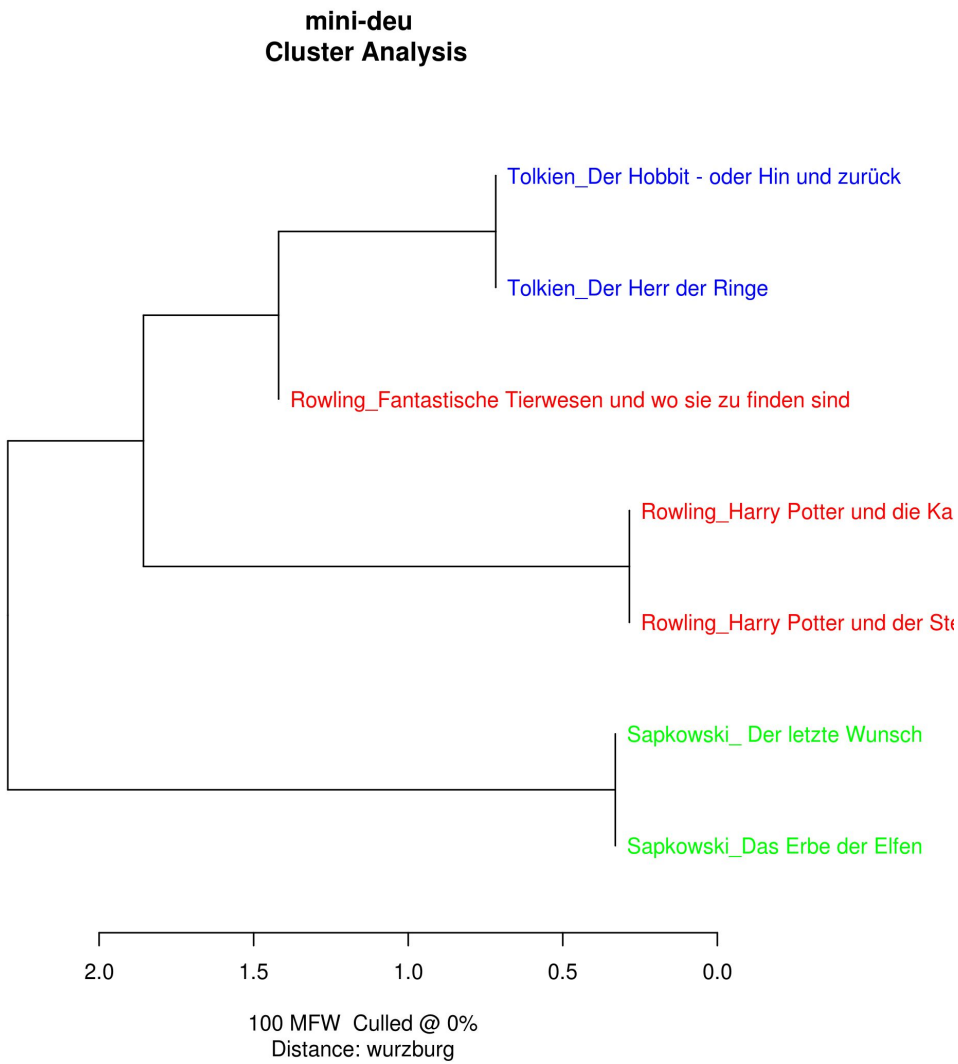
$f_x(T)$ = raw frequency of word *x* in text T;

μ_x = mean frequency of word *x* in a collection of texts;

σ_x = standard deviation of frequency of word *x*.

Classification algorithm

- k-nearest neighbors (v. good)
- Support Vector Machine (best)
- Nearest Shrunken Centroids
- Naive Bayes



Visualisation

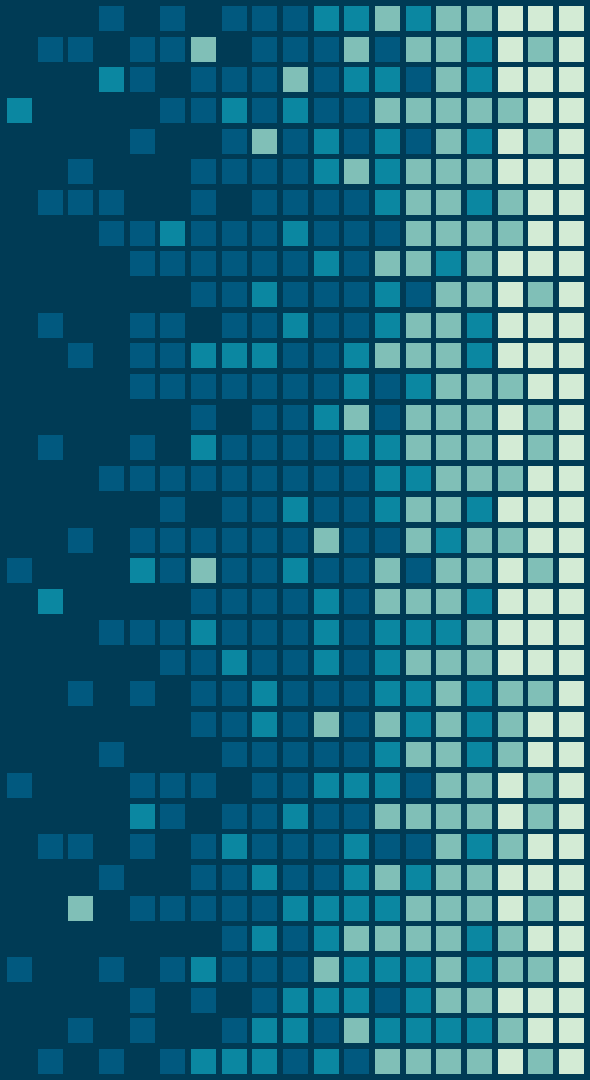
Hierarchical clustering analysis

(and many others)

Applications of stylometry

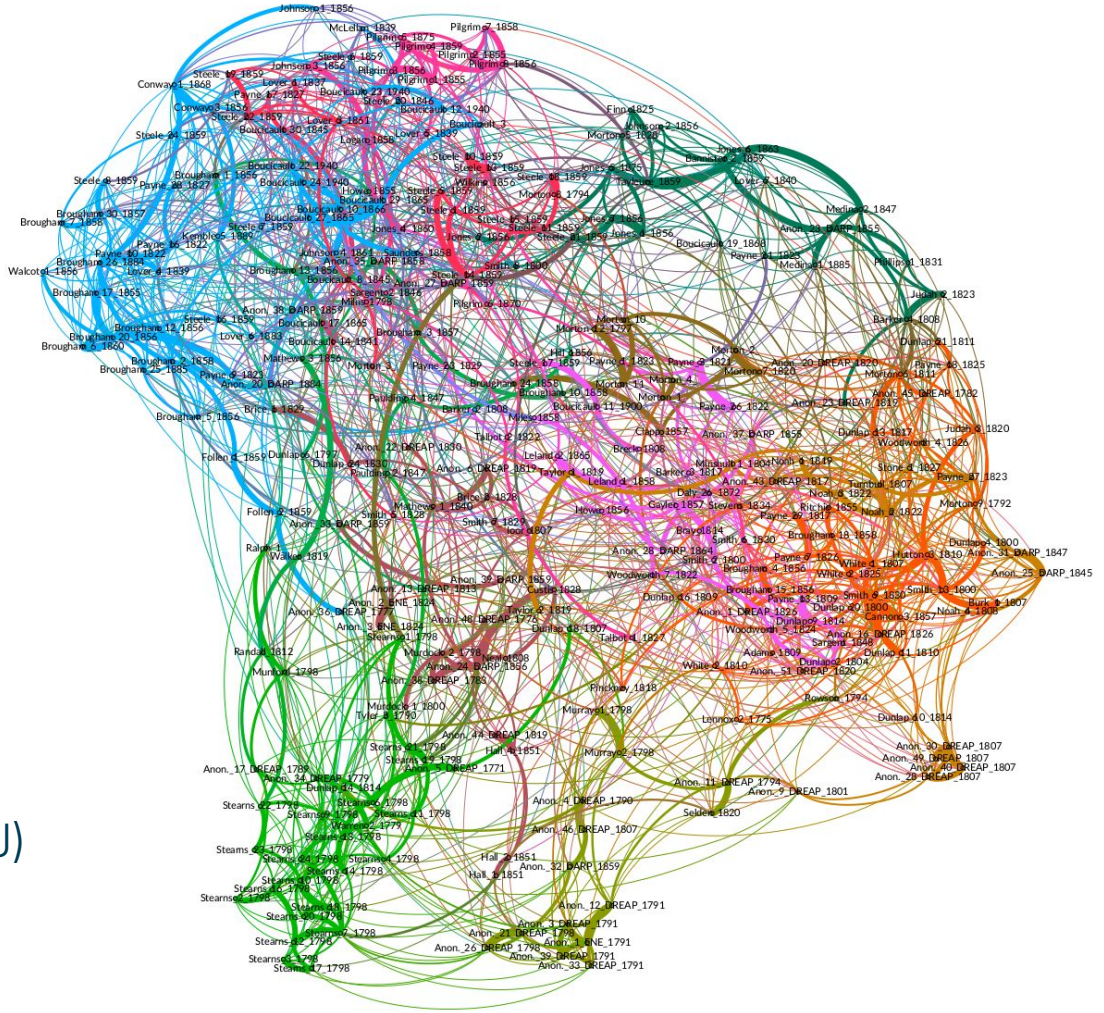
- authorship attribution,
- tracing chronology,
- analysis of cross and inter genre relationships,
- big data analysis,
- style transfer and anonymization,
- … and many others.

Stylometry in exploration



Making sense of a big American drama corpus

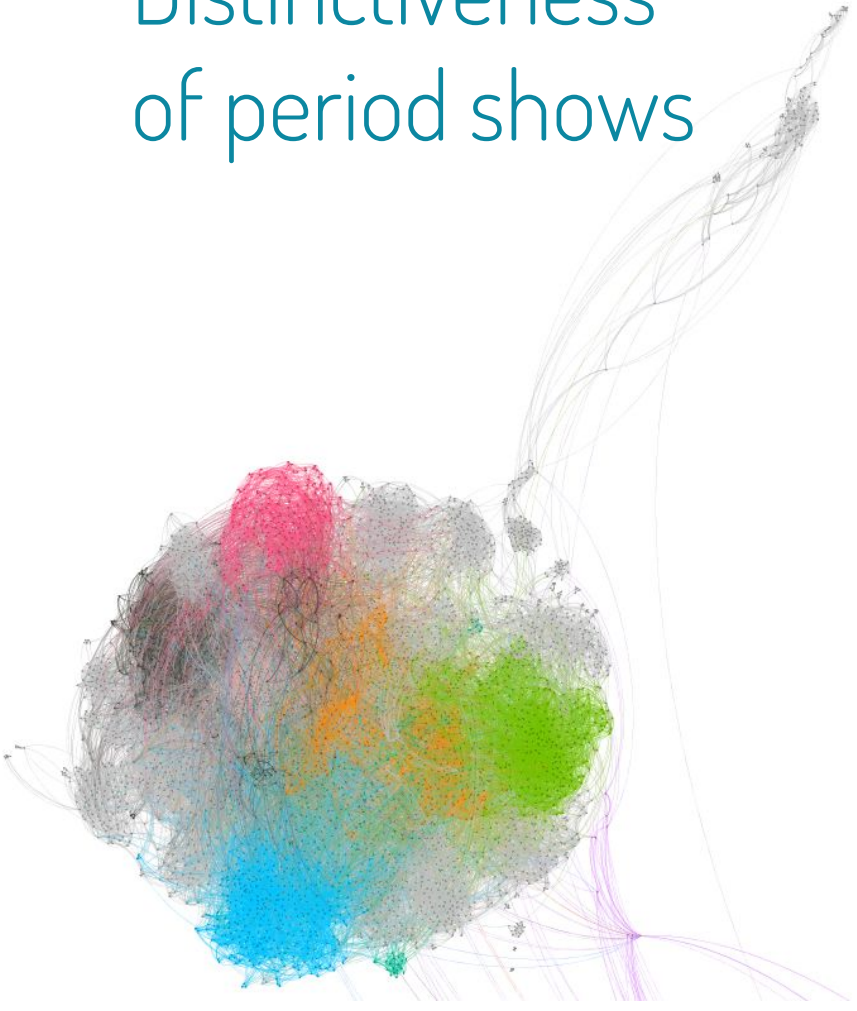
With Dennis Mischke (Potsdam U), Michał Choiński (Jagiellonian U)



Examining influences in TV series

- 90 British and American shows
- Mix of genres
- Genre? Writer? Producer?

Distinctiveness of period shows

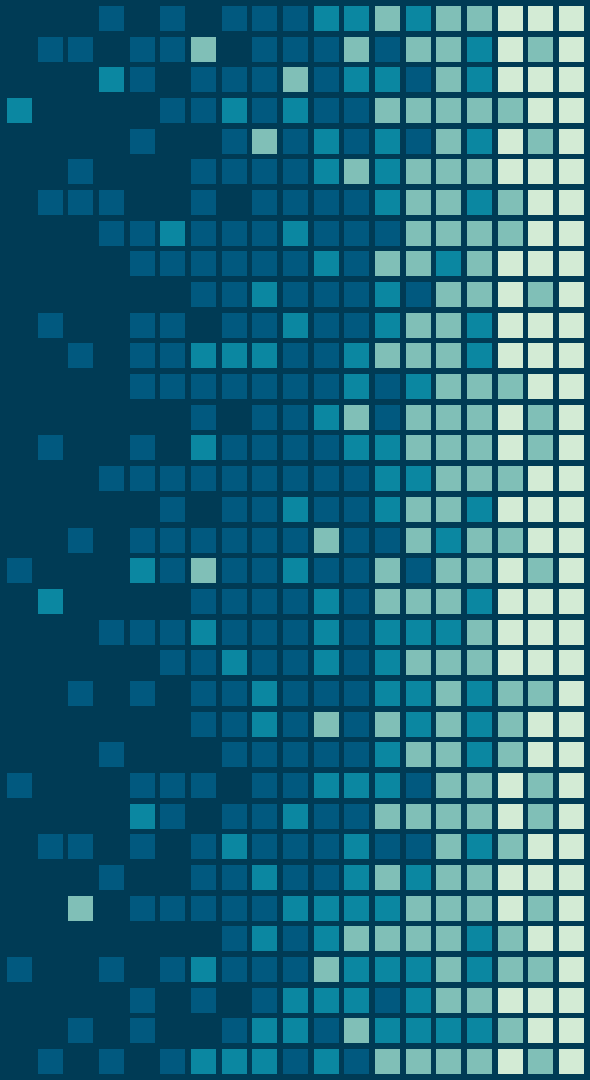


Period fiction:
The White Queen, The White Princess, The Tudors, Victoria, etc.

ABC + a little Showtime + BBC's Spooks, no common genre or topic

Production company?

Stylometry in authorship attribution



Classic authorship problems

Federalist papers, JK Rowling

Federalist Papers as an attribution case

- "A series of essays, anonymously published defending the document to the public"

(Lin-Manuel Miranda 2015)

- 85 texts authored by: Alexander Hamilton (51?), James Madison (29?) and John Jay (5)
- 12 letters of disputed authorship determined by stylometry

Mosteller, F., and D. L. Wallace (1964). Applied Bayesian and Classical Inference: The Case of The Federalist Papers. (and numerous other studies)

JK Rowling or Robert Galbraith?

- Who wrote “The Cuckoo's Calling”?
- Study by Patrick Juola (2013)

  “comparing against Rowling's own The Casual Vacancy, Ruth Rendell's The St. Zita Society, P.D. James' The Private Patient and Val McDermid's The Wire in the Blood…. Of the 11 sections of Cuckoo, six were closest (in distribution of word lengths) to Rowling, five to James.”

- Confirmed by the author

Language variation – idiolects

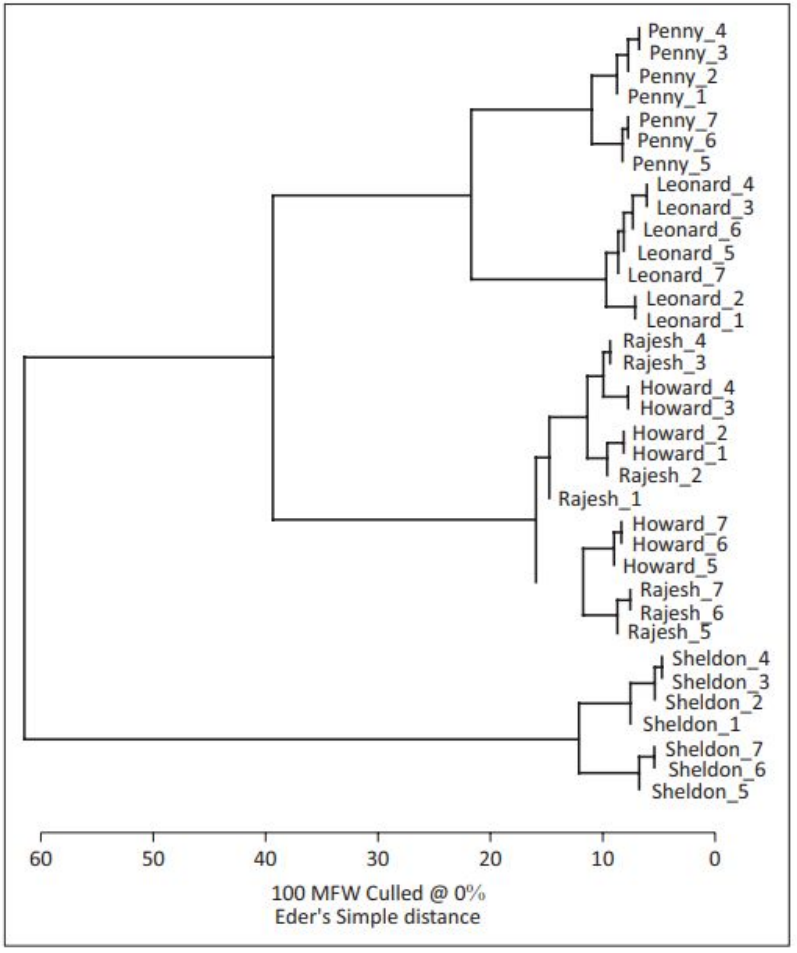
What is an idiolect?

An idiolect is the dialect of an individual person at one time. This term implies an awareness that no two persons speak in exactly the same way and that each person's dialect is constantly undergoing change—e.g., by the introduction of newly acquired words. Most recent investigations emphasize the versatility of each person's speech habits according to levels or styles of language usage.

Can we quantify character's idiolect?

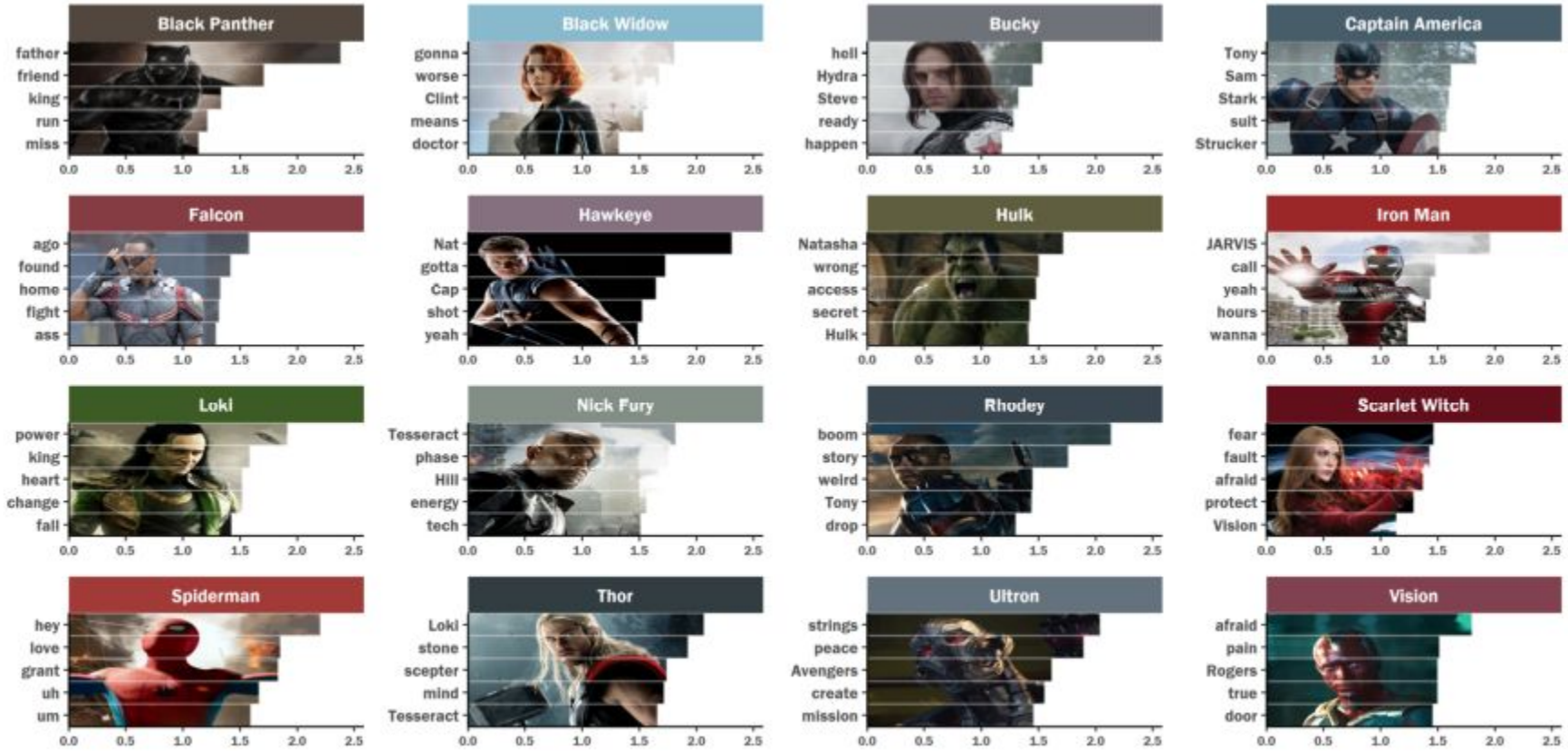
Source:

Van Zyl, M. & Botha, Y., 2016, 'Stylometry and characterisation in The Big Bang Theory', Literator 37(2), a1282 (https://literator.org.za/index.php/literator/article/view/1282/2148)



MFW, most frequent words.

FIGURE 1: Cluster analysis for *The Big Bang Theory*, Seasons 1–7.



https://towardsdatascience.com/i-analyzed-marvel-movie-scripts-to-learn-what-each-avenger-says-most-2e5e7b6105bf

The Voices of Doctor Who

Doctor Who (1963 –)

1963-89 ‘Classic’ series

- focus on the main character

1996 Film

2005- ? revival of the show: New/Nu series

- transition to authorial American-like model, increased role of a showrunner

Its main character, the Doctor, travels with his companions in the TARDIS (Time and Relative Dimension in Space), a ship capable of traveling through space and time that takes the exterior form of a 1930s British police booth, but is bigger on the inside. (...) the Doctor is not consistently portrayed by the same actor; periodically the Doctor “dies” and regenerates in a new humanlike form with a new personality [Edwards 2014, 375].

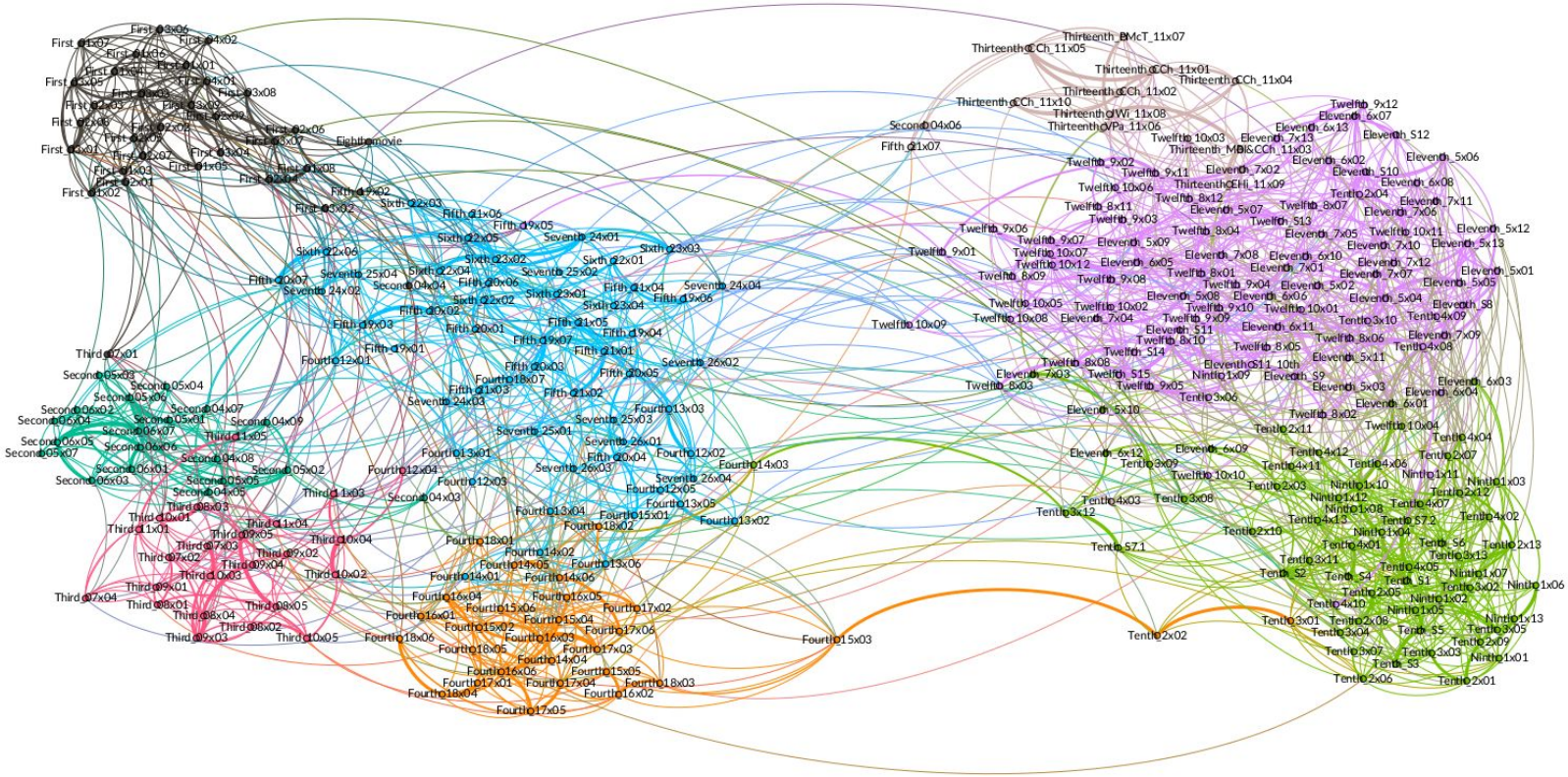
Methods

- A corpus of dialogue lines
- Network analysis
  - Bootstrap Consensus Tree in Stylo
  - Visualization in Gephi
  - Community Detection Algorithms (here Louvain's modularity algorithm)
- Rolling stylometry

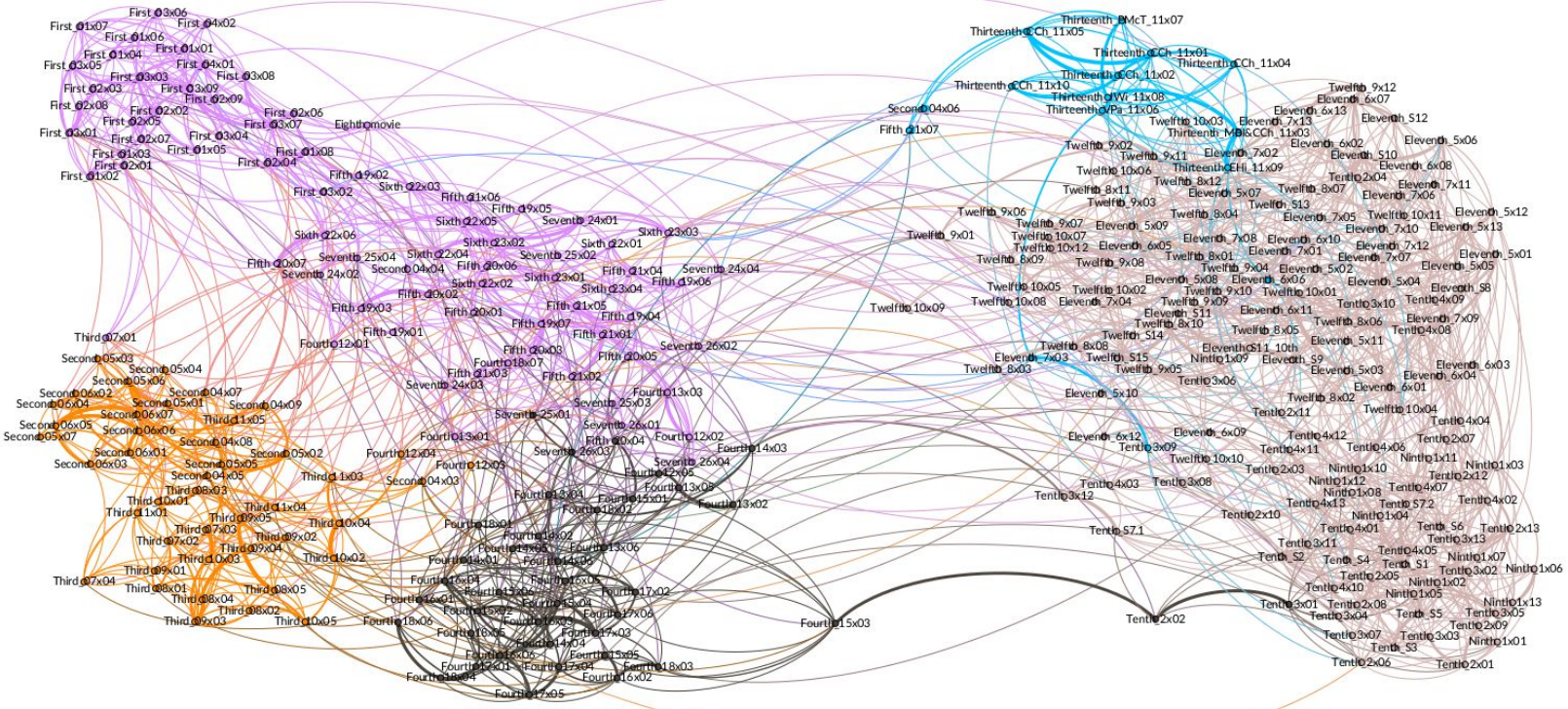
Selecting features

- High number of proper names – > tackled with ‘culling’ method
- Short texts – > 100-500 Most Frequent Words as features
- Cosine Delta classifier – proved the most reliable in recent studies

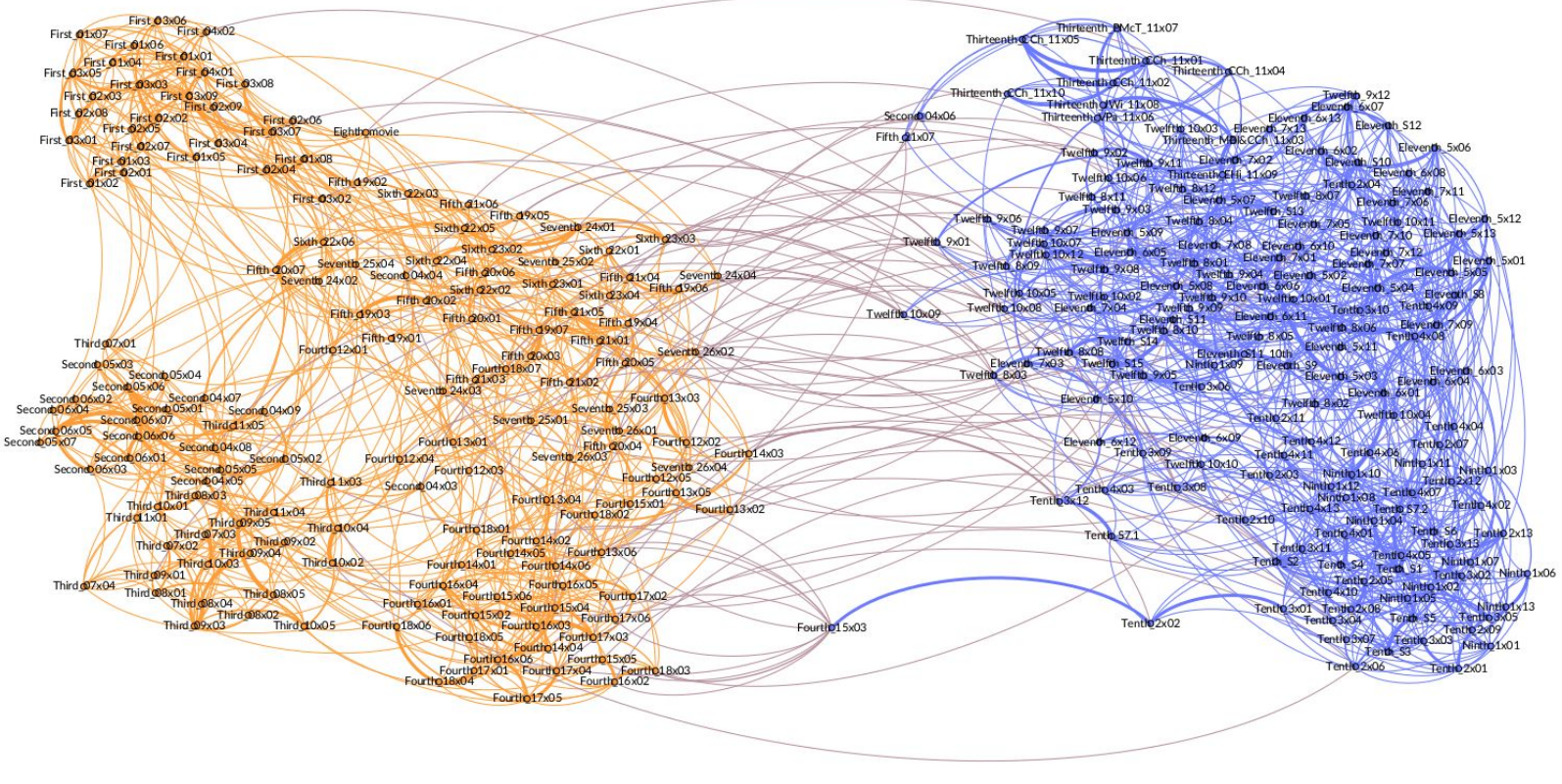
Just the Doctor (colored by regeneration=1)



Just the Doctor lines (colored by modularity=3)



Just the Doctor lines (colored by modularity=4)



Bonus: alienated First Doctor against regenerations 2-12, network analysis with modularity=5

Language variation – translators

Translators' voices

With examples from a study conducted with Quinn Dombrowski: 'Stylometric investigations into translationese: The Baby-Sitters Club across languages',

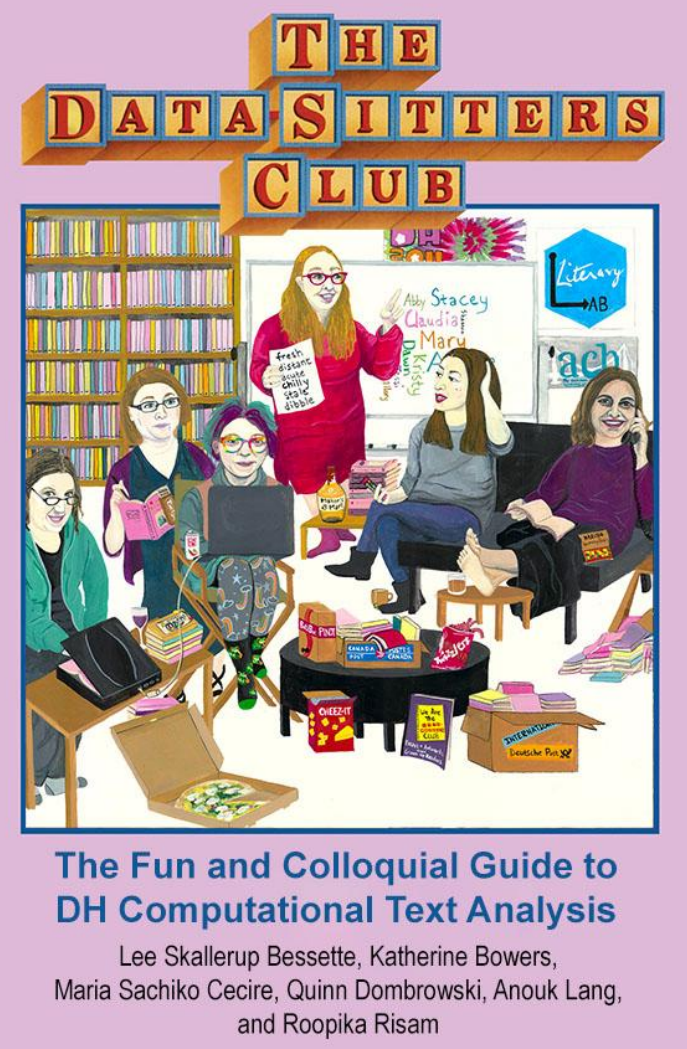
Translator's invisibility

A translated text (...) is judged acceptable by most publishers, reviewers, and readers when it reads fluently, (...) the appearance, in other words, that the translation is not in fact a translation, but the "original." (...) The more fluent the translation, the more invisible the translator, and, presumably, the more visible the writer or meaning of the foreign text.

(Venuti 1995: 1-2)

Our research question

- is the impact of ghostwriters stylometrically visible?
- do translators have visible style? (cf. Jan Rybicki on Virginia Woolf, discussion on Anita Raya being Elena Ferrante in *Drawing Elena Ferrante's Profile*, ed. A. Tuzzi, M.A. Cortelazzo)



Goals:

- apply DH methods to this corpus
- explain how they work in easy terms
- collect all the translations and compare them

Dataset: The Baby-Sitters Club

- a series of middle-grade novels written by Ann M. Martin, published from 1986 to 2000,
- translated into numerous languages, becoming international bestsellers

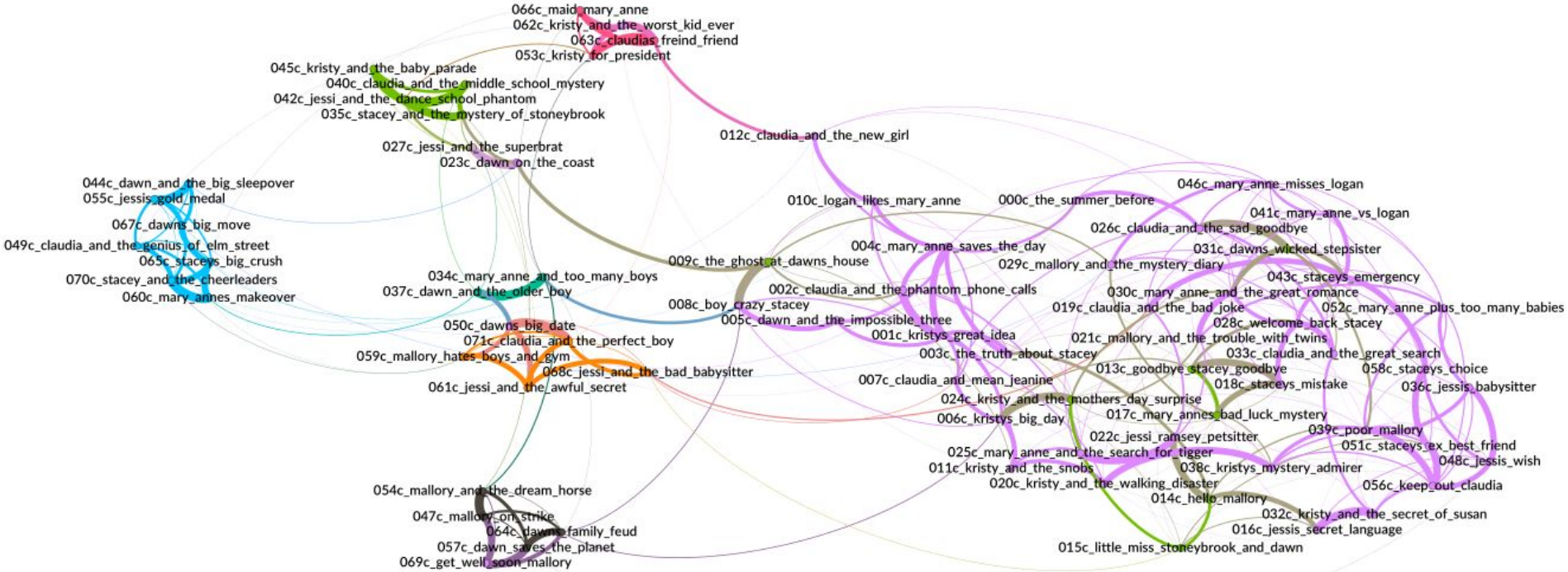
- our corpus: 142 translations into 6 language versions (distinguishing three French versions, next to Italian, Spanish, and Polish translations)

Our approach

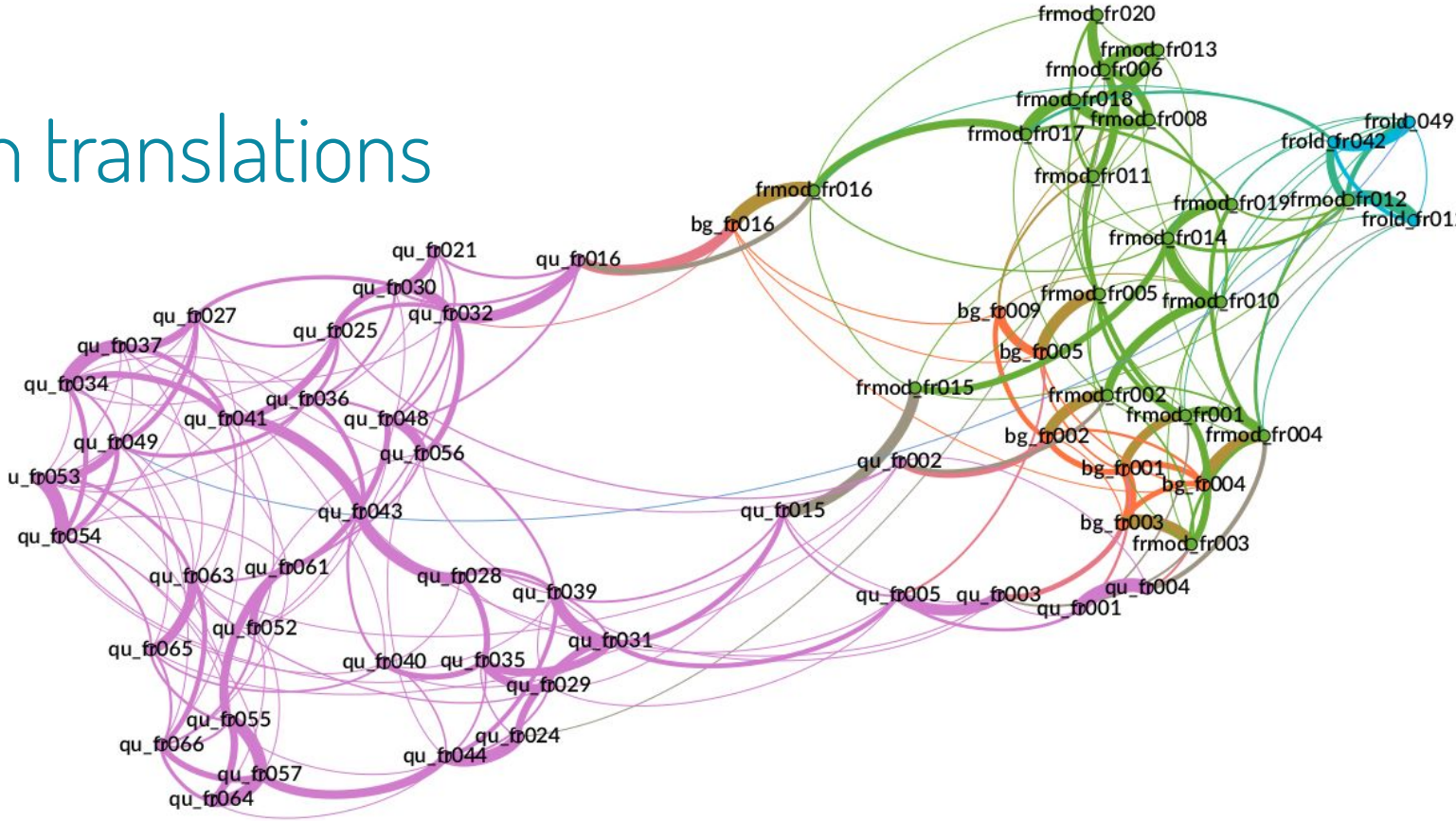
- cluster analysis
- bootstrap consensus tree + network visualization and additional modularity test
- 100-1000 MFW
- Cosine Delta and Burrows's Delta
- for EN and FR also tests with culling

1. English as a baseline for the relations between texts,
2. French, considering all three language variants together,
3. Italian,
4. Polish,
5. Spanish.

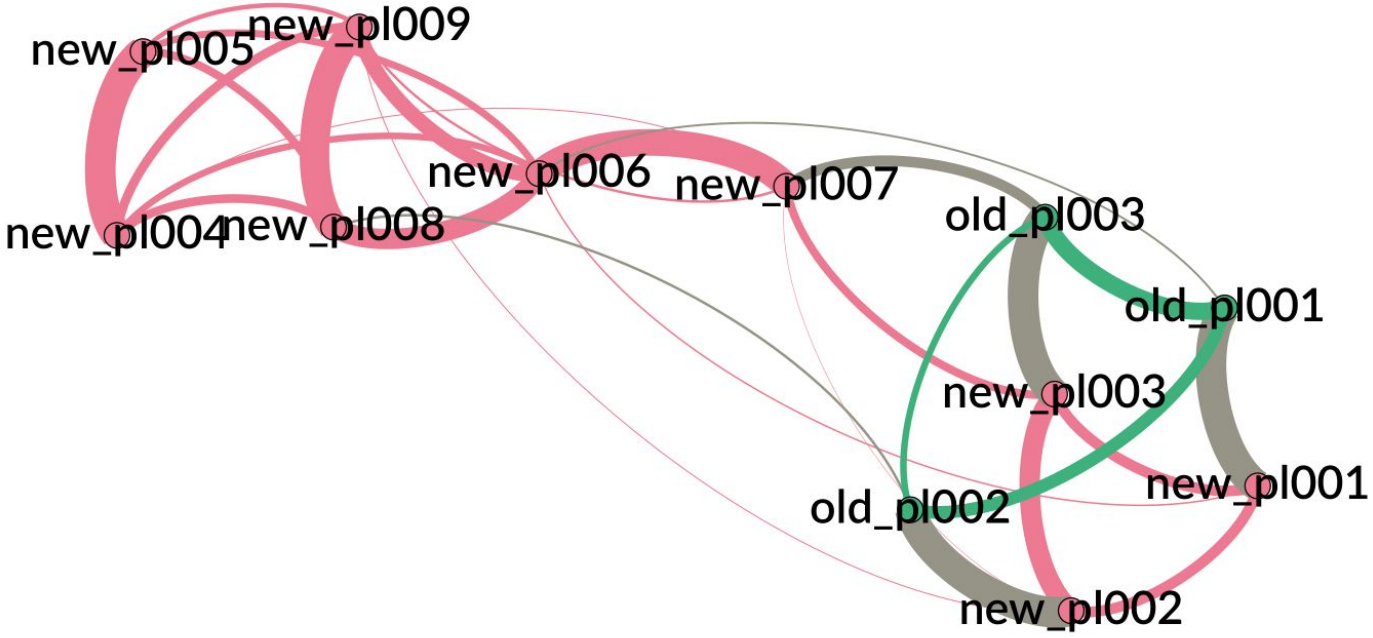
Originals (colors indicating (ghost –) writers



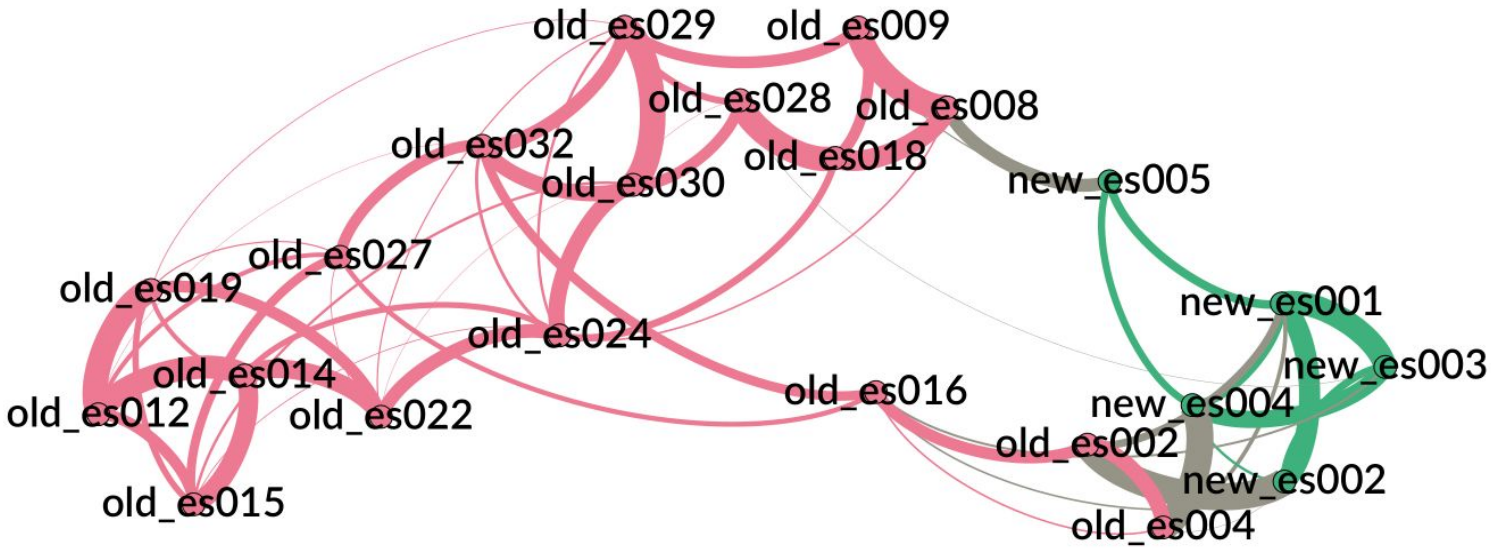
French translations



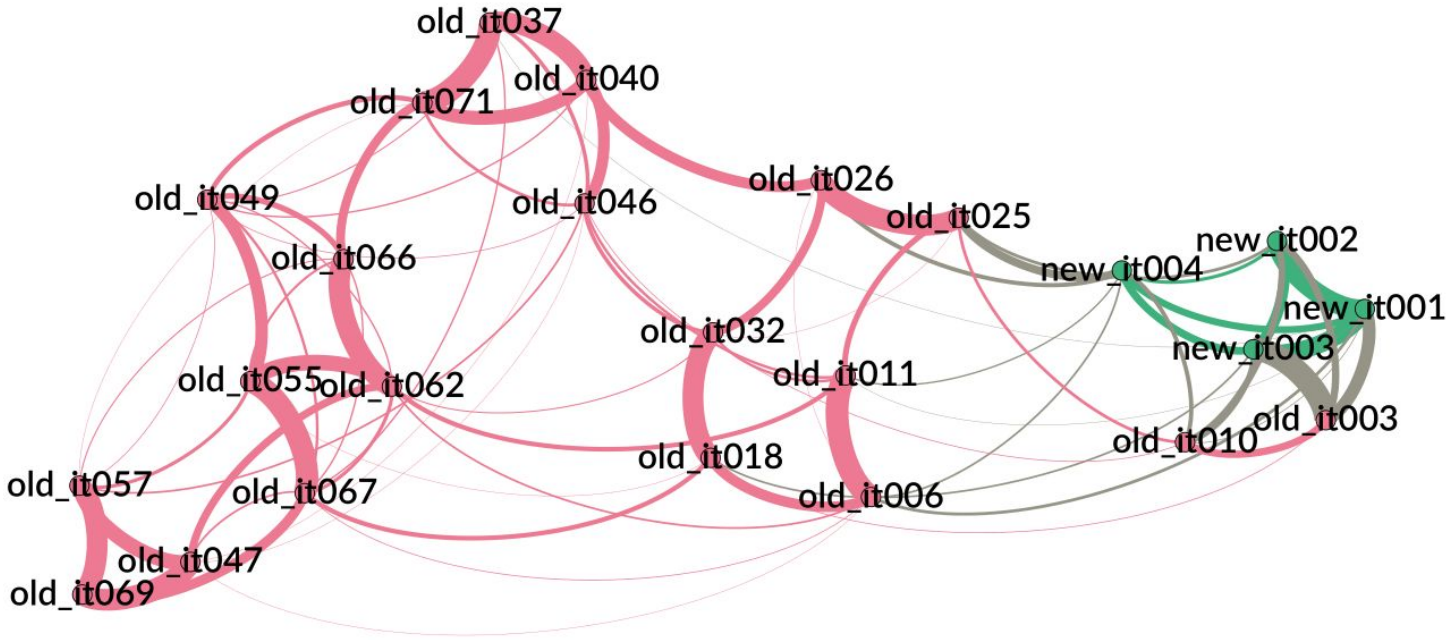
Polish translations



Spanish translations



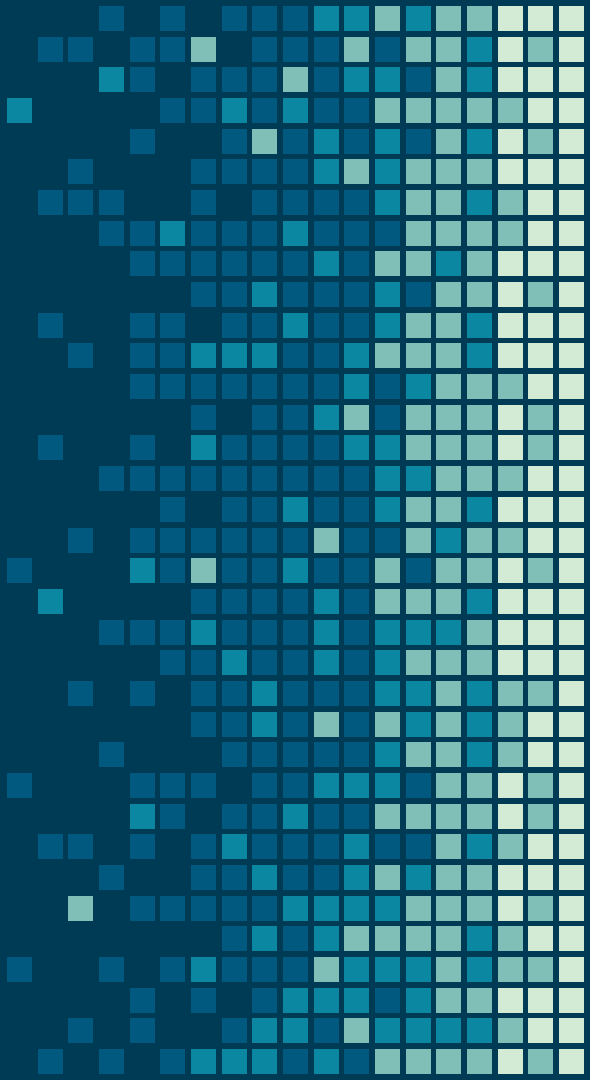
Italian translations



Conclusions

- some translators are more visible than others
- more so in some language-circles in our corpus = Spanish and Italian
- language variant carries strong stylistic signal
- ghostwriter is as visible as original author, also in the translation

Stylometry beyond text



Measuring style in dance

M. Escobar Varela and L. Hernández-Barraza. 'Digital Dance Scholarship: Biomechanics and culturally-situated dance analysis' in Digital Scholarship in the Humanities (2019).

Questions:

- use the biomechanical toolkit to address questions relevant to dance scholars
- identify the biomechanical markers of different character types for male dancers in the dramatic Sendratari form of Yogyakarta

Measuring style in dance

The best discriminator of humble versus proud qualities is the left hip (on the x plane), where higher ROM correlates with a humble quality and a lower ROM correlates with a proud quality.

Escobar Varela, M., Hernández-Barraza, L. „Digital Dance Scholarship: Biomechanics and Culturally Situated Dance Analysis". Digital Scholarship in the Humanities. https://doi.org/10.1093/llc/fqy083.

Screenshot my own from: https://villaorlado.github.io/dance/html/compareall.html

Stylometry of literary papyri

Ochab J.K., Essler H. **Stylometry of literary papyri**. In: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage - DATeCH2019. ACM Press; 2019. p. 139–42. Available from: http://dl.acm.org/citation.cfm?doid=3322905.3322930

Goals:

- authorship attribution
- automatic genre classification

And for documentary papyri:

- automatic extraction of formulaic expressions
- automatic genre classification
- supplementation of missing metadata
- enhancement of metadata and annotation

Stylometry of papyri

Ochab J.K., Essler H. **Stylometry of literary papyri**. In: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage - DATeCH2019. ACM Press; 2019. p. 139–42. Available from: http://dl.acm.org/citation.cfm?doid=3322905.3322930

Study:

- 298 texts from Digital Corpus of Literary Papyrology (DCLP).
- The metadata from the Leuven Database of Ancient Books (LDAB)
- 66 authors

Stylometry of papyri

Ochab J.K., Essler H. **Stylometry of literary papyri**. In: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage - DATeCH2019. ACM Press; 2019. p. 139–42. Available from: http://dl.acm.org/citation.cfm?doid=3322905.3322930

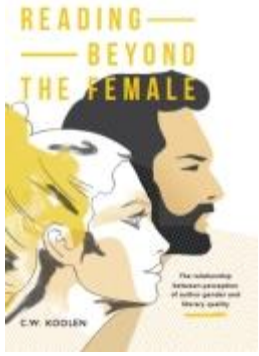
Findings:

- successful classification correlating text regularization (scribes' impact?)
- good chances of automatic genre distinction

Stylometry of literary quality

C.W. Koolen – Reading Beyond The Female

What is the relation between gender and perceived literary quality?



And many more...

- Comic books stylometry (Alexander Dunst)
- Music stylometry (e.g. Andrew Brinkman)
- Cinemetrics
- Multimodal stylometry?

Let’s try to do that!

https://computationalstylistics.github.io/stylo_nutshell/#main-functions-stylo

Getting started

Set working directory:

Command line:

```
setwd("the/path/to/my/favourite/folder")
```

RStudio users: find your directory in the Files panel, then use *Menu > More > Set as Working Directory*

Windows users: use *Menu > File > Change directory*