CLSINFRA COMPUTATIONAL LITERARY STUDIES INFRASTRUCTURE

Examining language variation with stylometry

Joanna Byszuk (Institute of Polish Language, Polish Academy of Sciences)
joanna.byszuk@ijp.pan.pl

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004984

Plan of the workshop

CLSINFRA

1. What is stylometry? What is language variation? – revision
2. How to use stylo – introduction to basic functions
3. Comparing idiolects with stylometry
4. Comparing translators with stylometry

What is stylometry?

What is stylometry?

CLSINFRA

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

How does it work?

CLSINFRA

corpus of texts

+

distance measure

+

classification algorithm

+

(visualisation)

What words at the top? What relevance?

CLSINFRA

- 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)

Stylometry is related to

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

Applications of stylometry

CLSINFRA

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

Stylometry beyond text

Measuring style in dance

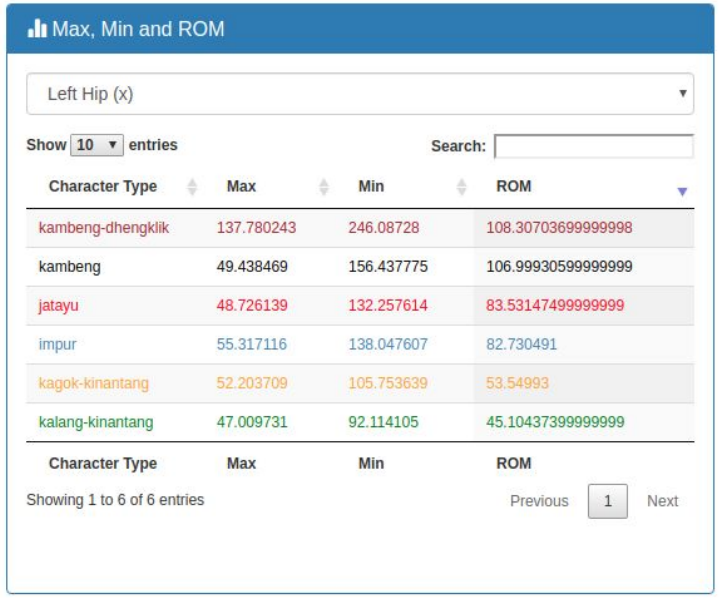
M. Escobar Varela and L. Hernández-Barraza. 'Digital Dance Scholarship: Biomechanics and culturally-situated dance analysis' in Digital Scholarship in the Humanities (2019).

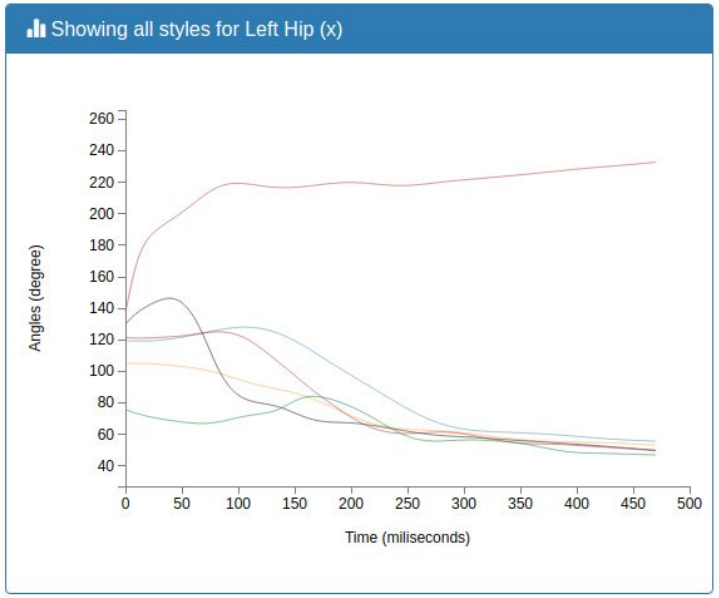
Questions:

- 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

Comparison Matrix (all styles for one joint)





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

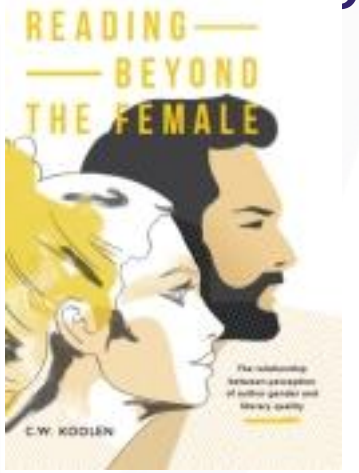
Findings:

- 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?

Language variation – idiolects

What is an idiolect?

CLSINFRA

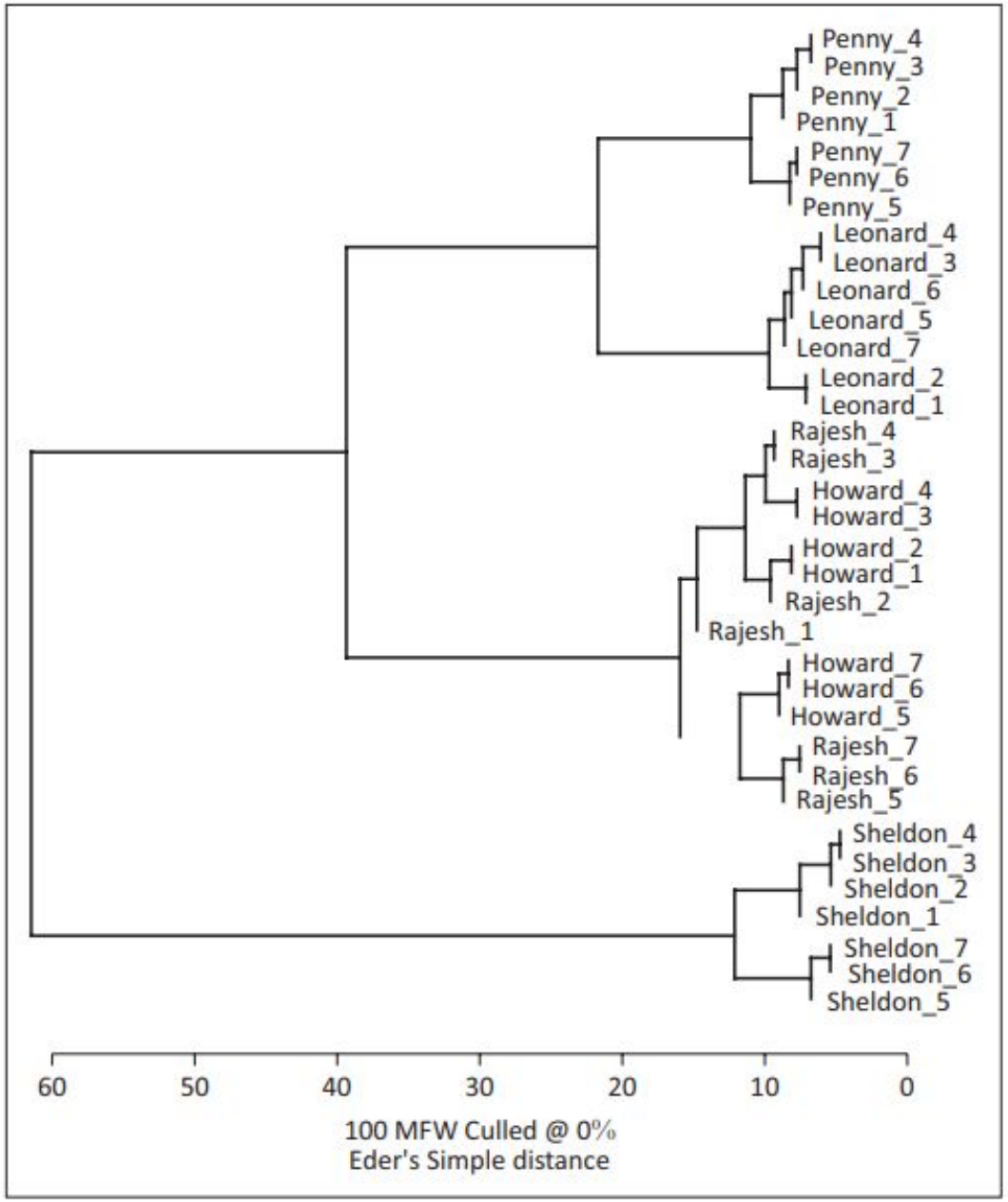
An idiolect is the dialect of an individual person at one time. This term implies an awareness that no two persons speak in exactly the same way and that each person's dialect is constantly undergoing change—e.g., by the introduction of newly acquired words. Most recent investigations emphasize the versatility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

Source: https://www.britannica.com/topic/dialect#ref1046707

Can we quantify character's idiolect?

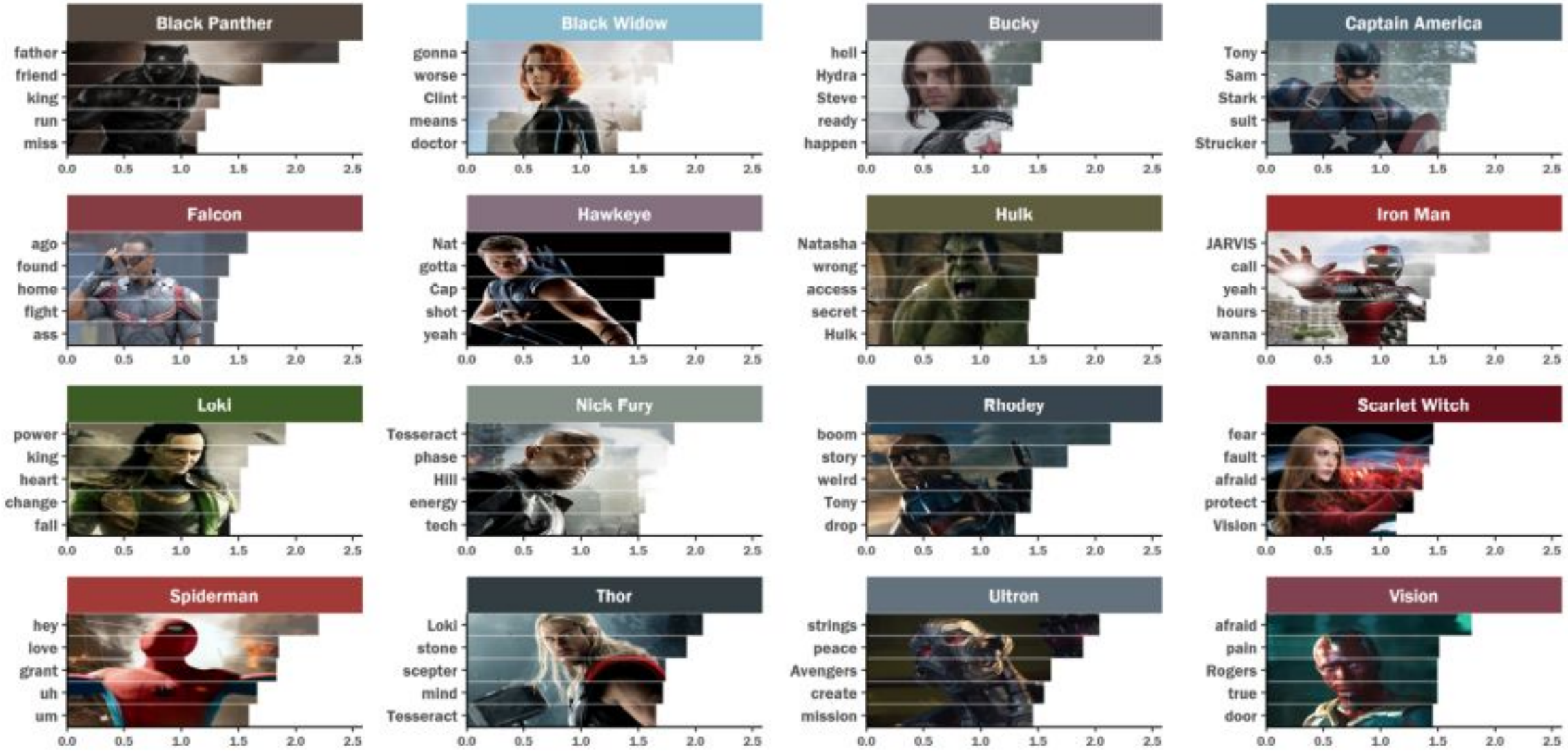
Source:

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 –)

CLSINFRA

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].

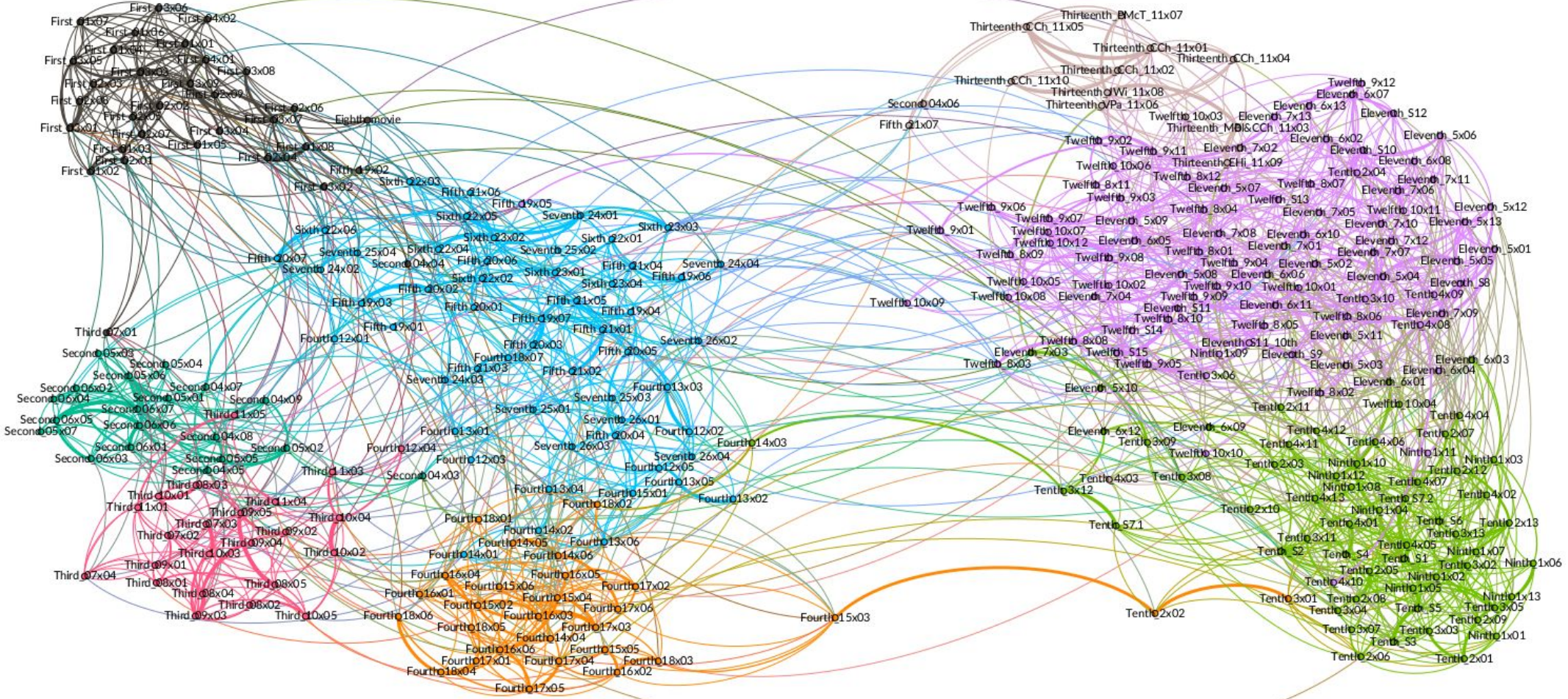
Methods

- 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

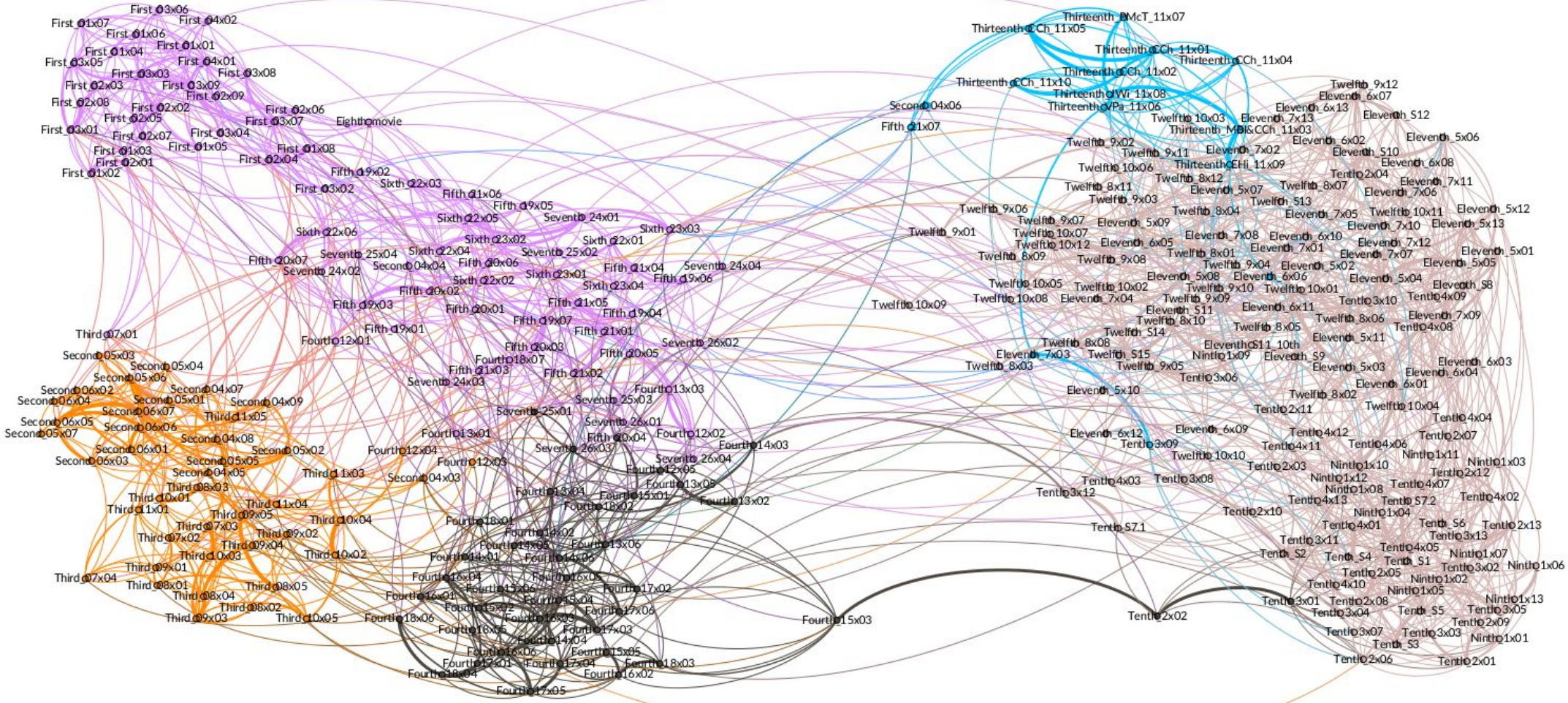
Selecting features

- 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

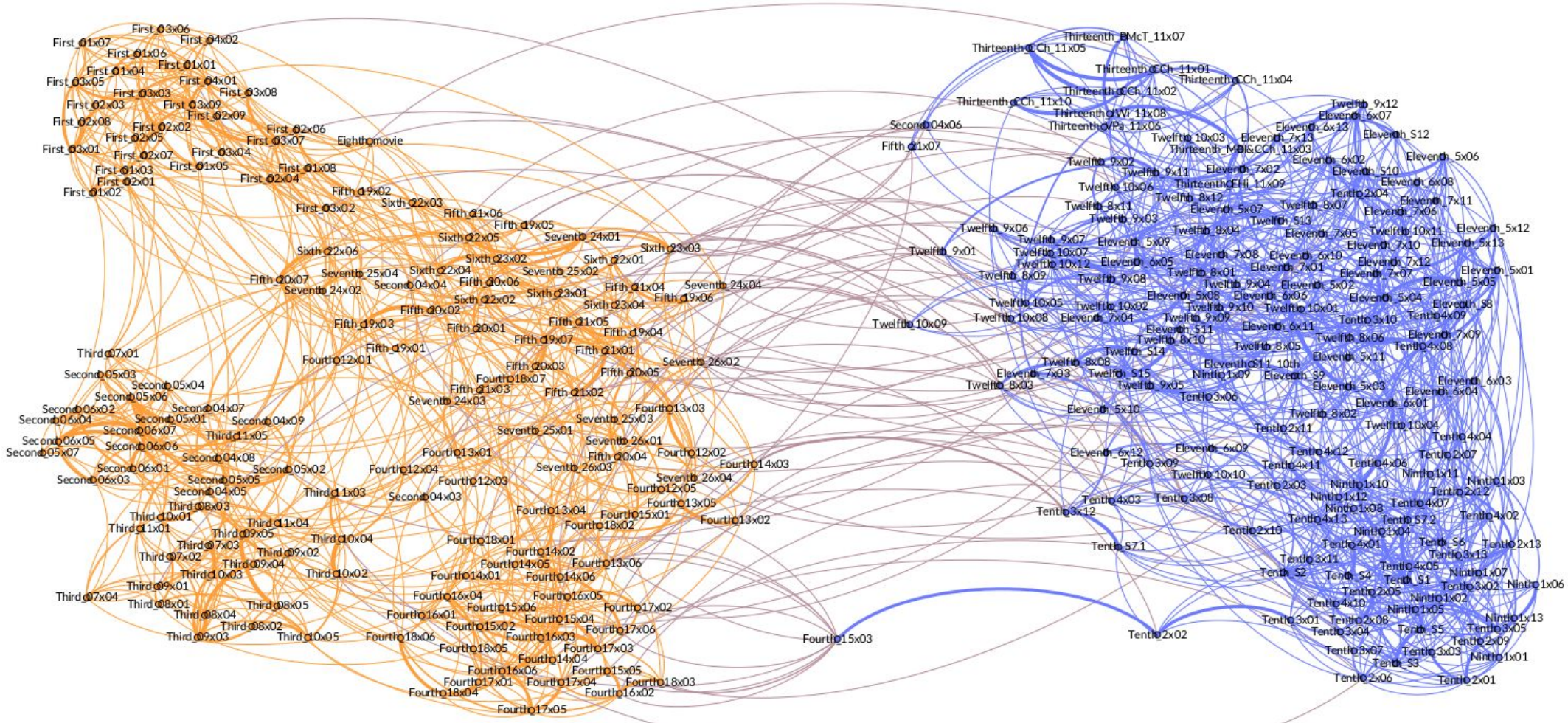
Just the Doctor (colored by regeneration=1)



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

CLSINFRA

With examples from the studies: “What Makes a Captain: Quantitative Analysis of Discourses of Power across Star Trek Series” 2022, and

Language in TV idiolects

fake spoken language - usually lacking narration element explaining what's happening -> less context and related ... words, more I-you phrases, more phrases for keeping the conversation flowing - emotive, conative and phatic language functions perhaps more present than referential and poetic - but also other functions as in Bednarek

Functions of TV language

- highly stylized and conventionalized to resemble “natural” dialogue.
- allows for slower pace and more filler dialogue,
- aimed to sound "natural", but the utterances are usually shorter and differ in the use of various linguistic features, so it can:
  - transfer explicit or implicit thematic messages,
  - create continuity and consistency, or
  - provide more insight into the character.

Bednarek (2018: 35-77).

Star Trek

CLSINFRA

- a series of sci-fi series and films, including 8 live-action and 3 animated series as well as 13 films
- a cultural phenomenon since 1966, with cult following and tons of amateur and professional research
- height of popularity reflecting in development in the 80-90s, and again since 2017

Why study Star Trek?

CLSINFRA

- considered *progressive* and formative sci-fi
- reflecting changes in social expectations for the future
- consider groundbreaking when it comes to inclusion and diversity:
  - 1966 TOS – Asian and Black actors as regulars
  - 1993 DS9 – first Black captain
  - 1995 VOY – first female captain
  - 2017 Discovery – first female Asian and later Black captains, regular LGBTQ characters

Starfleet command order

With the exception c
chain of command

– >

insight into how the
characters and genc

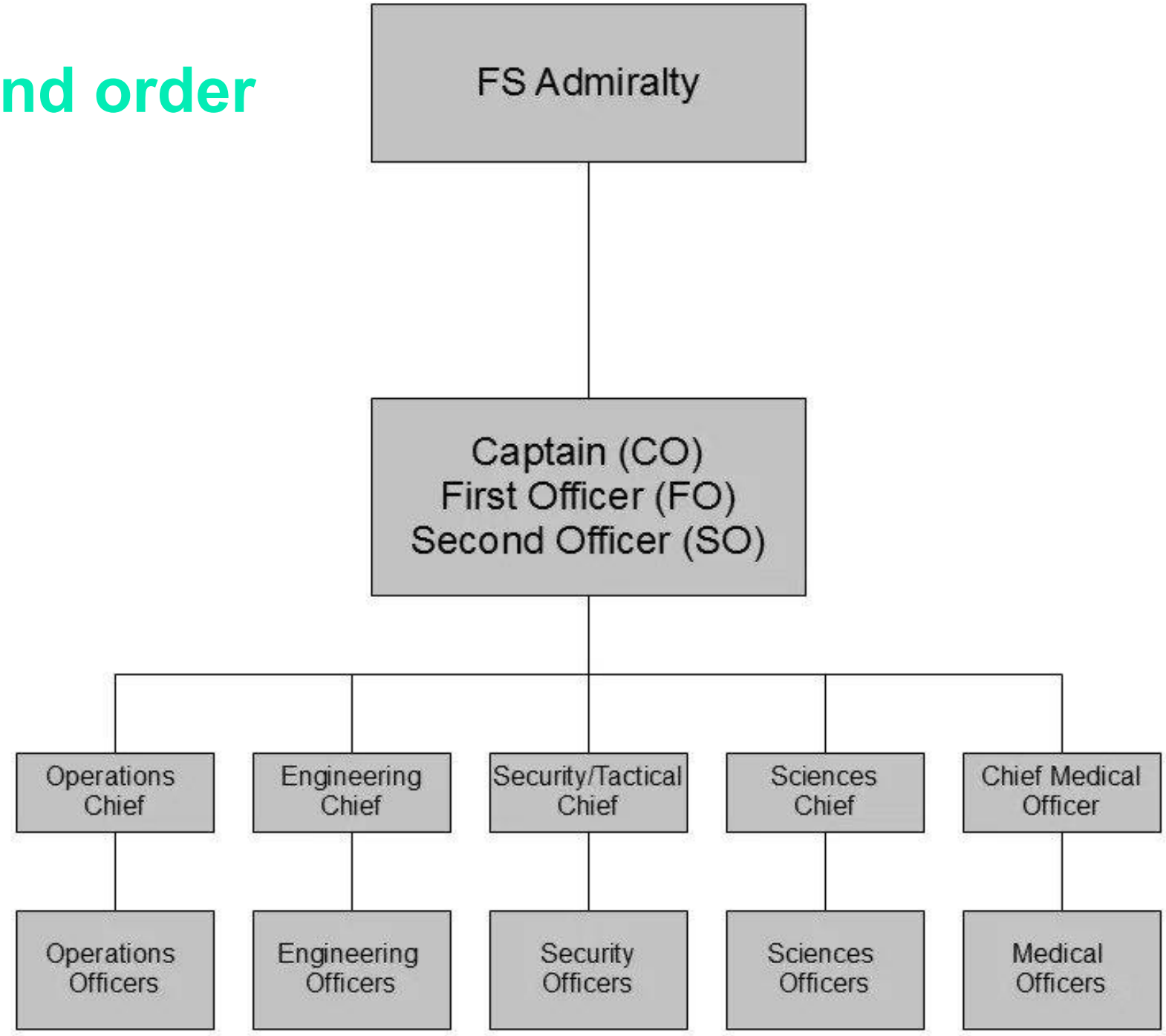
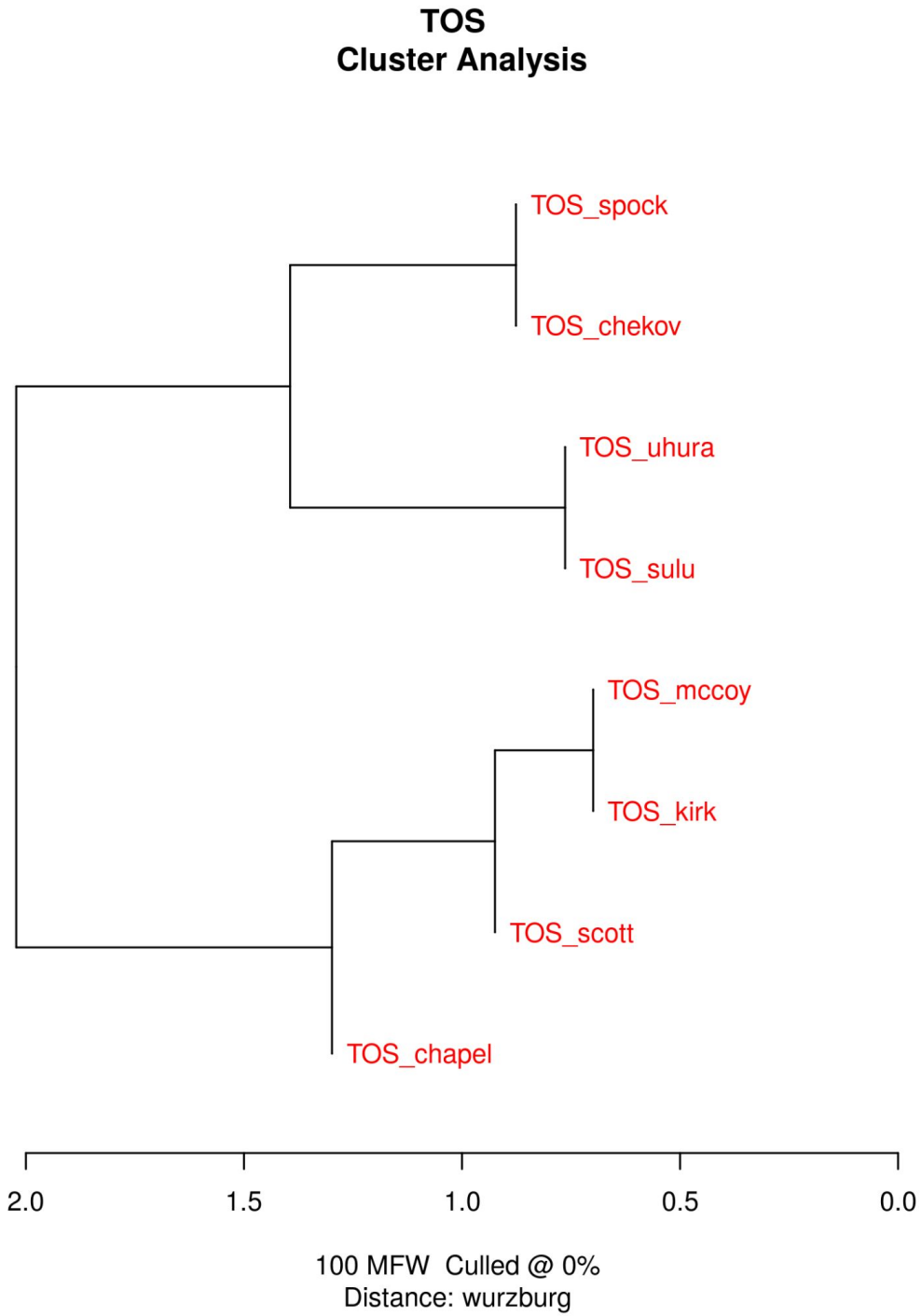
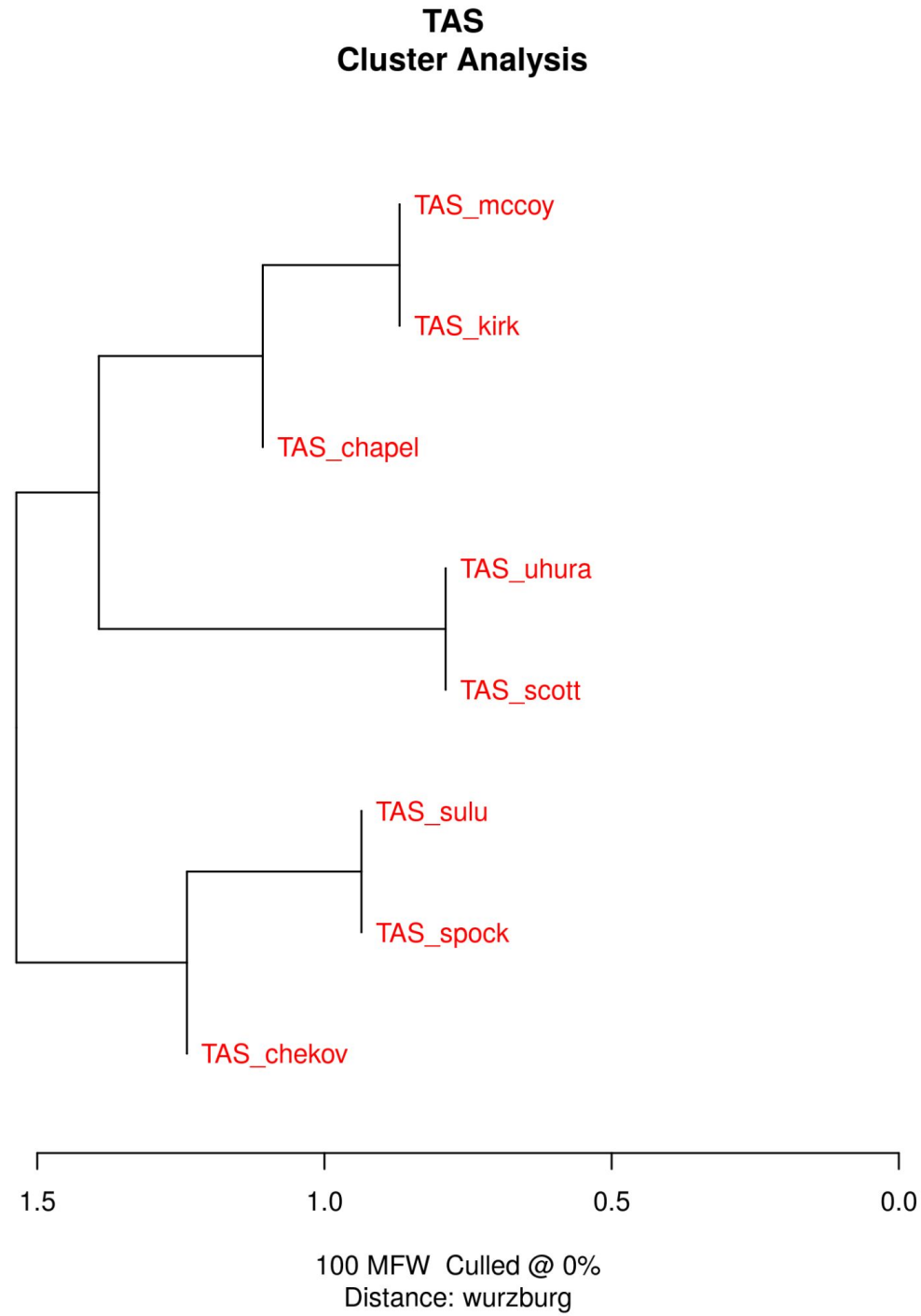


Image source:
https://federation-starfleet.fandom.com/wiki/Chain_of_Command

TOS / TAS (1966-68/ 73-74)

Mr. Spock (80)
James T. Kirk (79)
Leonard McCoy (76)
[Nyota] Uhura (67)
Montgomery Scott (65)
Hikaru Sulu (51)
Pavel Chekov (36)
Nurse [Christine] Chapel (28)
Lt. Kyle (11)
Janice Rand (8)

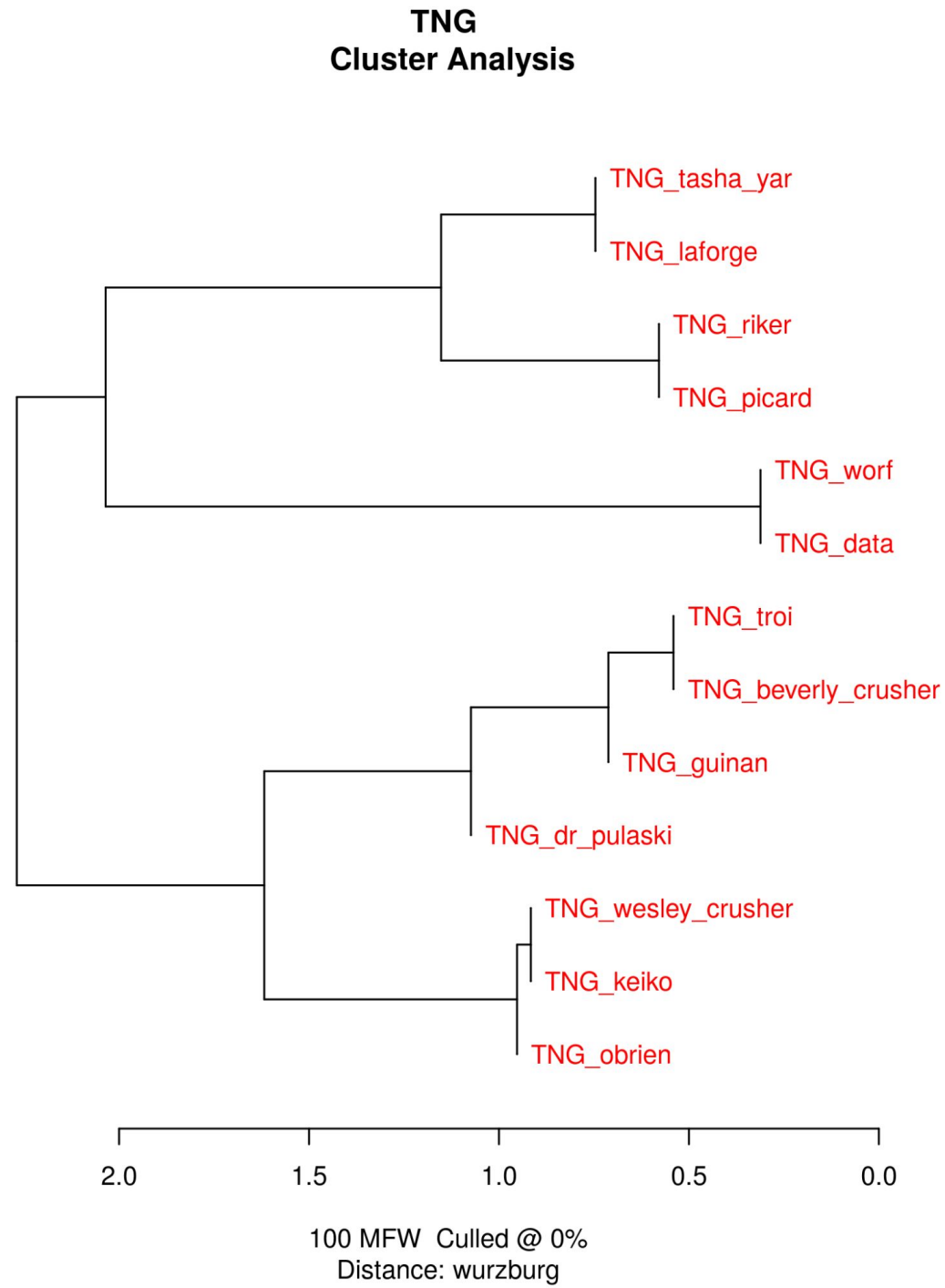

TOS
Cluster Analysis
TOS_spock
TOS_chekov
TOS_uhura
TOS_sulu
TOS_mccoy
TOS_kirk
TOS_scott
TOS_chapel
2.0
1.5
1.0
0.5
0.0
100 MFW Culled @ 0%
Distance: wurzburg


TAS
Cluster Analysis
TAS_mccoy
TAS_kirk
TAS_chapel
TAS_uhura
TAS_scott
TAS_sulu
TAS_spock
TAS_chekov
1.5
1.0
0.5
0.0
100 MFW Culled @ 0%
Distance: wurzburg

TNG (1987-1994)

Capt. Jean-Luc Picard (176)
Cmdr. William Riker (176)
Lt. Cmdr. Geordi La Forge (176)
Couns. Deanna Troi (176)
Lt. Cmdr. Data (176)
Lt. Worf (176)
Dr Beverly Crusher (154)
Chief Miles O'Brien (82)

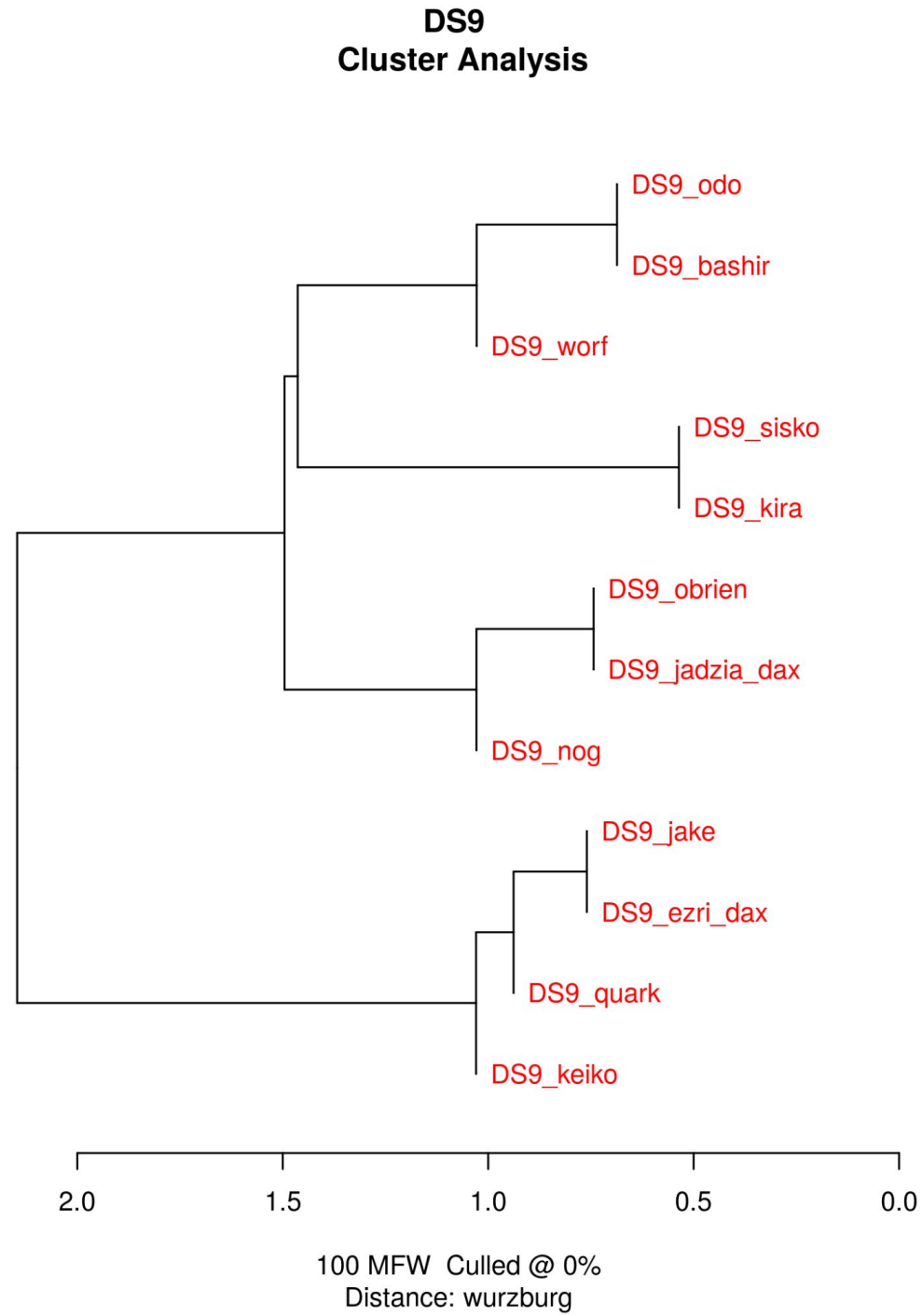
Guinan (28)
Lt. Tasha Yar (28)
Wesley Crusher (86)
Dr Pulaski (20)
Nurse Ogawa (16)
Ensign Ro Laren (9)
Keiko O'Brien (8)


TNG
Cluster Analysis
TNG_tasha_yar
TNG_laforge
TNG_riker
TNG_picard
TNG_worf
TNG_data
TNG_troi
TNG_beverly_crusher
TNG_guinan
TNG_dr_pulaski
TNG_wesley_crusher
TNG_keiko
TNG_obrien
2.0
1.5
1.0
0.5
0.0
100 MFW Culled @ 0%
Distance: wurzburg

DS9 (1993-99)

Capt. Benjamin Sisko (173)
Odo (173)
Doctor Bashir (173)
Chief Miles O'Brien (173)
Major Kira (173)
Quark (173)
Lt. Cmdr. Jadzia Dax (148)
Lt. Cmdr. Worf (102)
Lt. Ezri Dax (25)

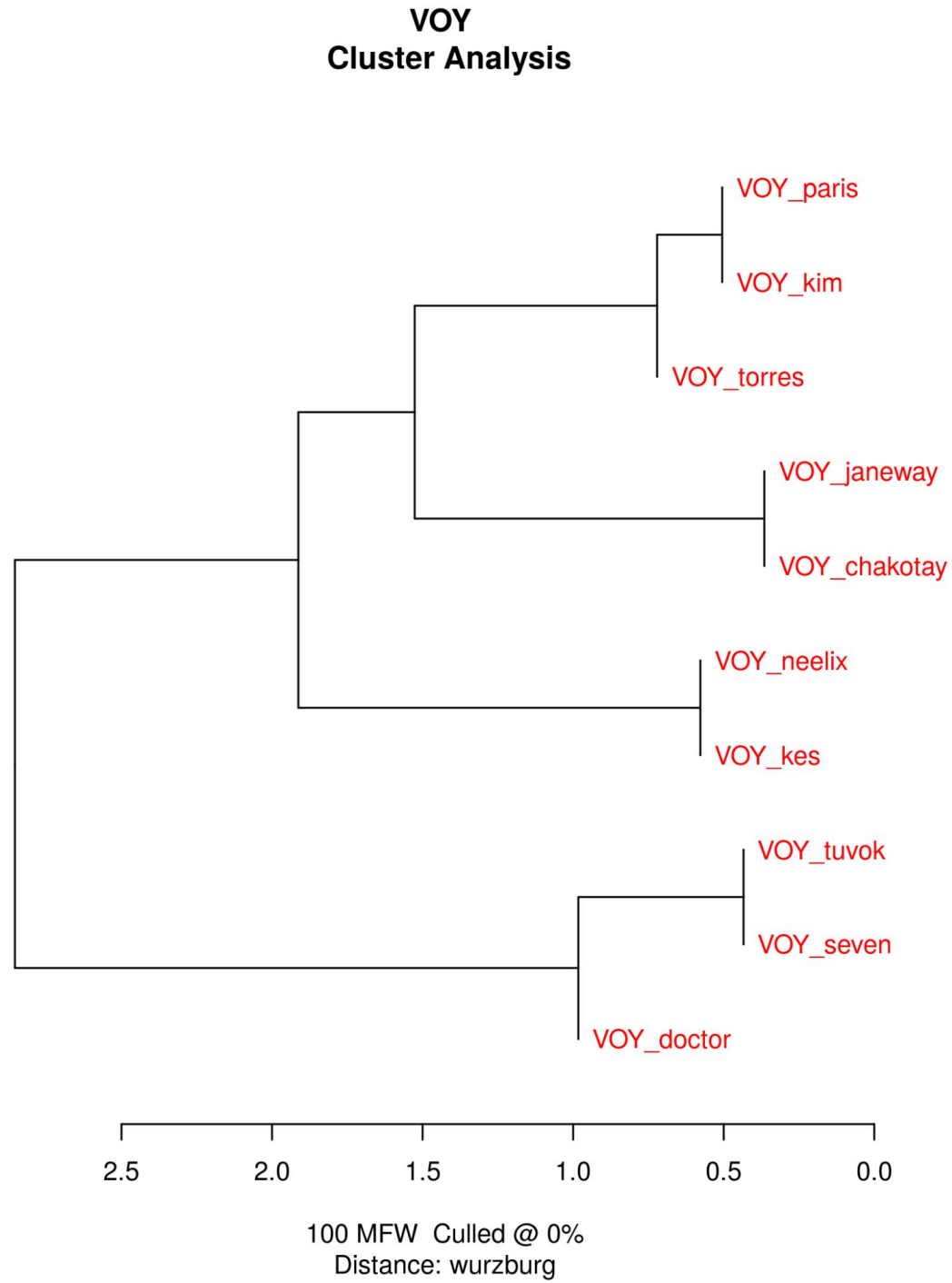
Jake Sisko (173)
Nog (47)
Keiko O'Brien (19)


DS9
Cluster Analysis
DS9_odo
DS9_bashir
DS9_worf
DS9_sisko
DS9_kira
DS9_obrien
DS9_jadzia_dax
DS9_nog
DS9_jake
DS9_ezri_dax
DS9_quark
DS9_keiko
2.0
1.5
1.0
0.5
0.0
100 MFW Culled @ 0%
Distance: wurzburg

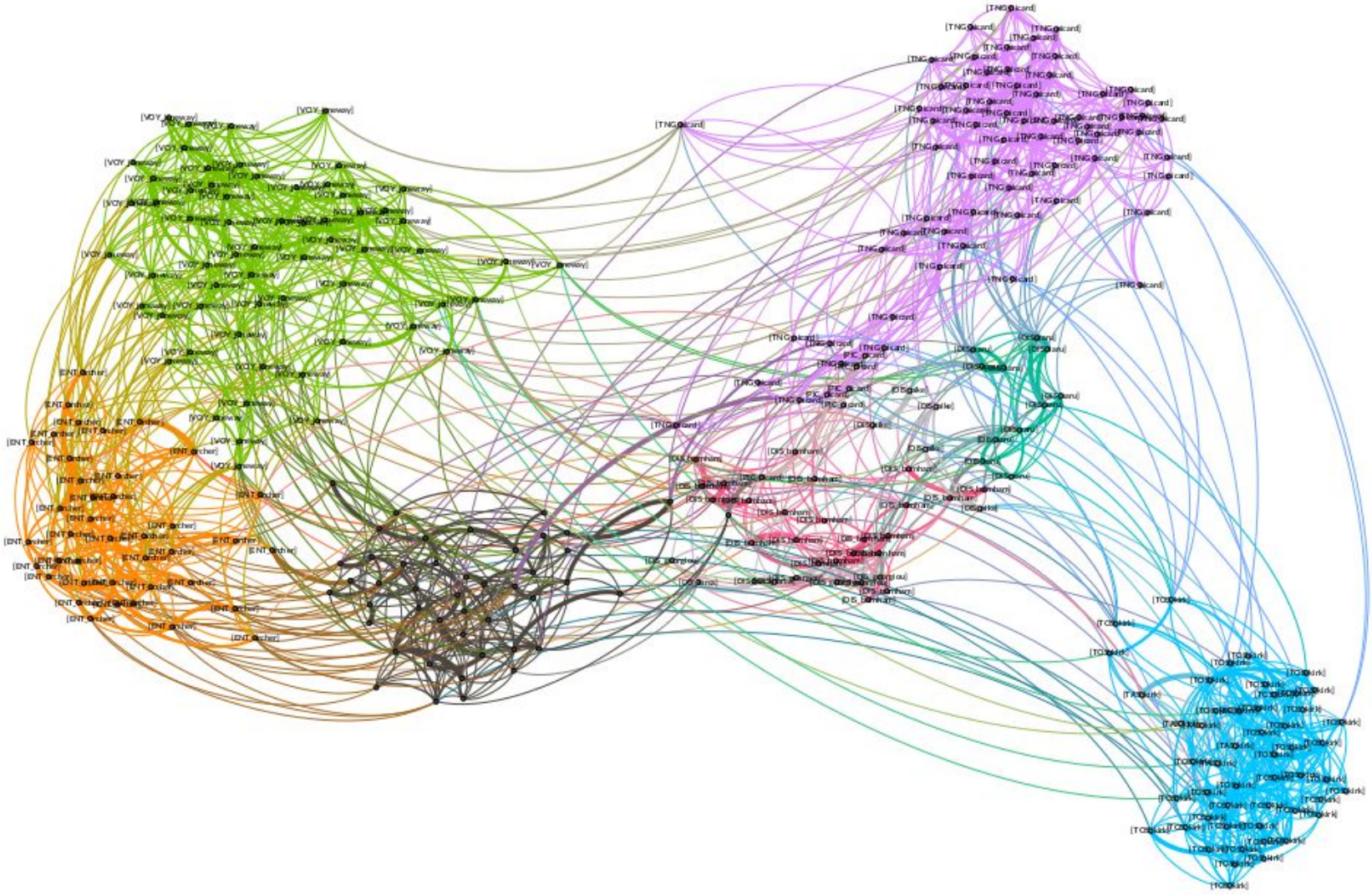
VOY (1995-2001)

Capt. Kathryn Janeway (168)
Cmdr. Chakotay (168)
Lt. B'Elanna Torres (168)
Lt. Tom Paris (168)
Neelix (168)
The Doctor (168)
Lt. Tuvok (168)
Ens. Harry Kim (168)

Seven of Nine (101)
Kes (70)


VOY
Cluster Analysis
VOY_paris
VOY_kim
VOY_torres
VOY_janeway
VOY_chakotay
VOY_neelix
VOY_kes
VOY_tuvok
VOY_seven
VOY_doctor
2.5
2.0
1.5
1.0
0.5
0.0
100 MFW Culled @ 0%
Distance: wurzburg

Captains

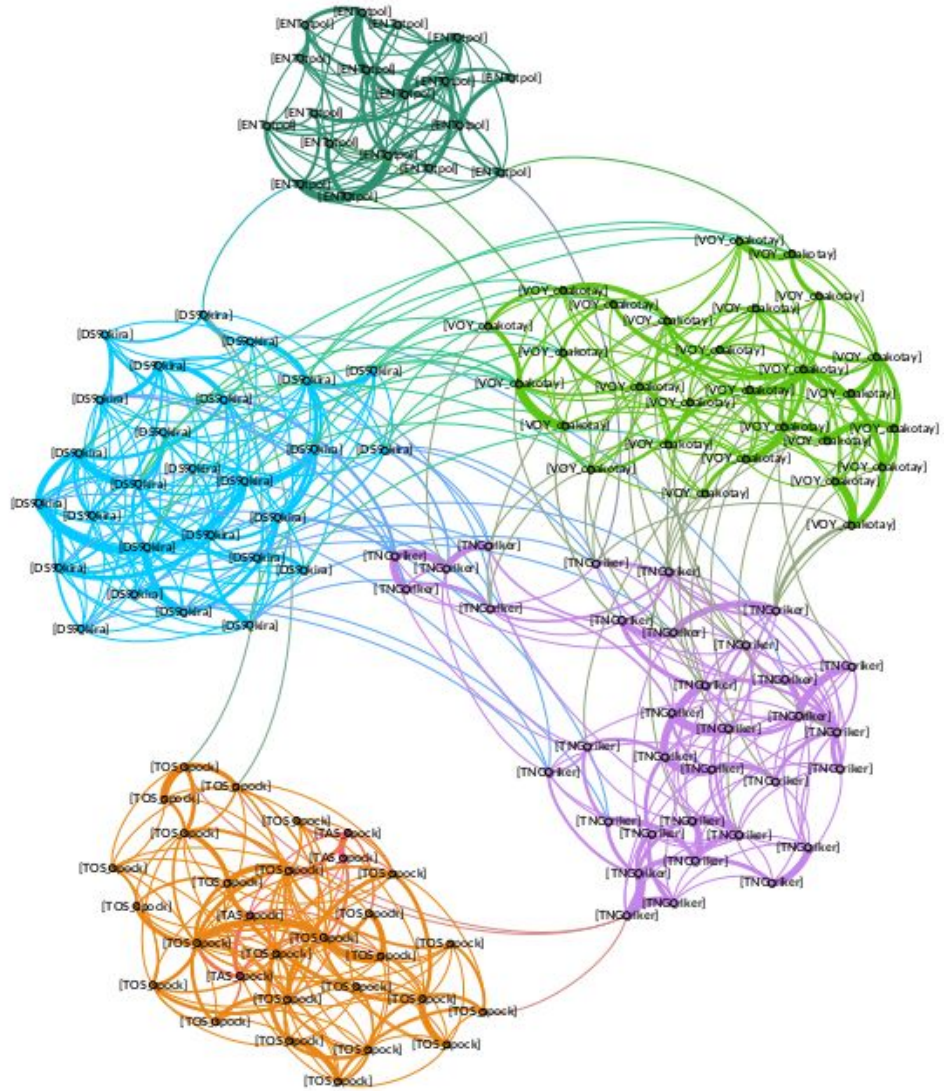


Captains – top collocations

what do you
i want to
what is it
i don^t know
we have to
i want you
out of here
this is the
be able to
want you to
we^re going to
as soon as
i^m going to

what are you
us out of
i don^t think
to the bridge
do you know
you want to
not going to
this is captain
do you think
a lot of
i have to
on my way
do you have
i^d like to

First Officers

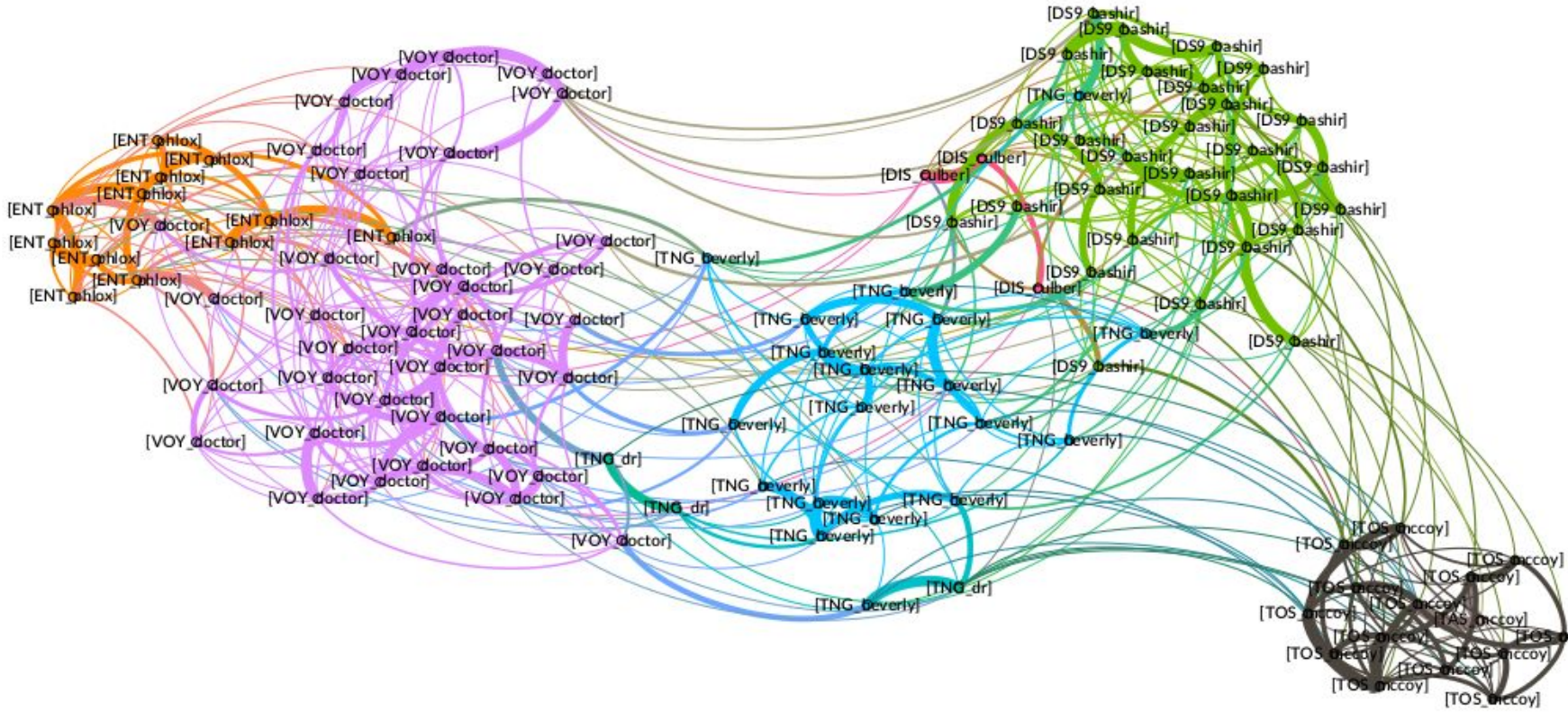


First officers – top collocations

be able to
i don^t know
what do you
to the bridge
going to be
what are you
one of the
i don^t think
out of here
we^re going to
some kind of
this is the
i^m going to

a lot of
do you think
we have to
what is it
to be a
i want to
i have to
i^d like to
going to have
i do not
i^m not sure

Doctors


[VOY doctor]
[ENT phlox]
[TNG dr]
[TNG beverly]
[DS9 bashir]
[DIS culber]
[TOS mccoy]

Doctors – top collocations

i don^t know
be able to
i^m going to
i^d like to
what are you
i don^t think
what do you
going to be
you^re going to
how do you
i^m not sure
i need to
do you think

i have to
we have to
what is it
you have a
i can do
one of the
i want to
some kind of
i don^t understand
a matter of
i have a
this is the
want you to

Conclusions

- series signal stronger than function
- TOS/TAS most different from other series
- otherness > gender
- captains = more orders, other staff = more doubts and questions

Language variation – translators

Translators' voices

With examples from a study conducted with Quinn Dombrowski: 'Stylometric investigations into translationese: The Baby-Sitters Club across languages', 2022 JADT conference

'

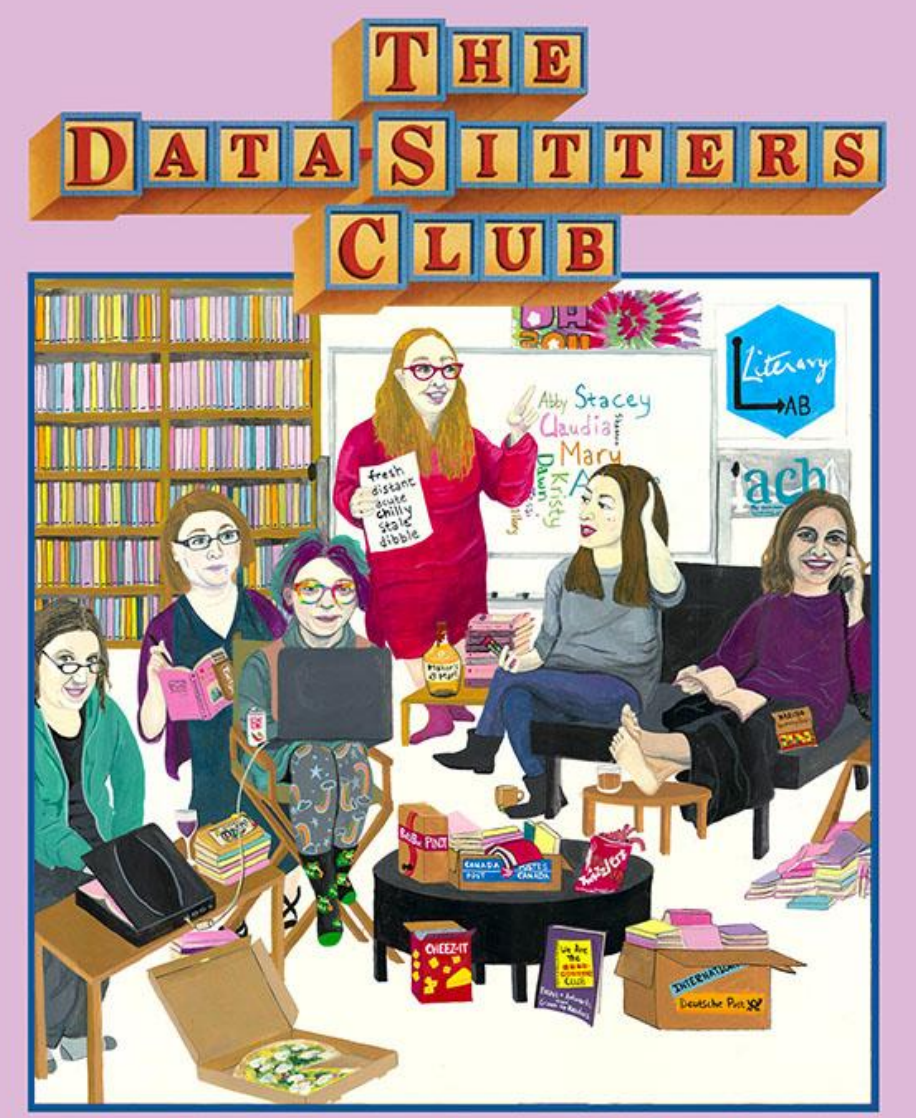
Translator's invisibility

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)



CLSINFRA

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

CLSINFRA

- 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)

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)

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.

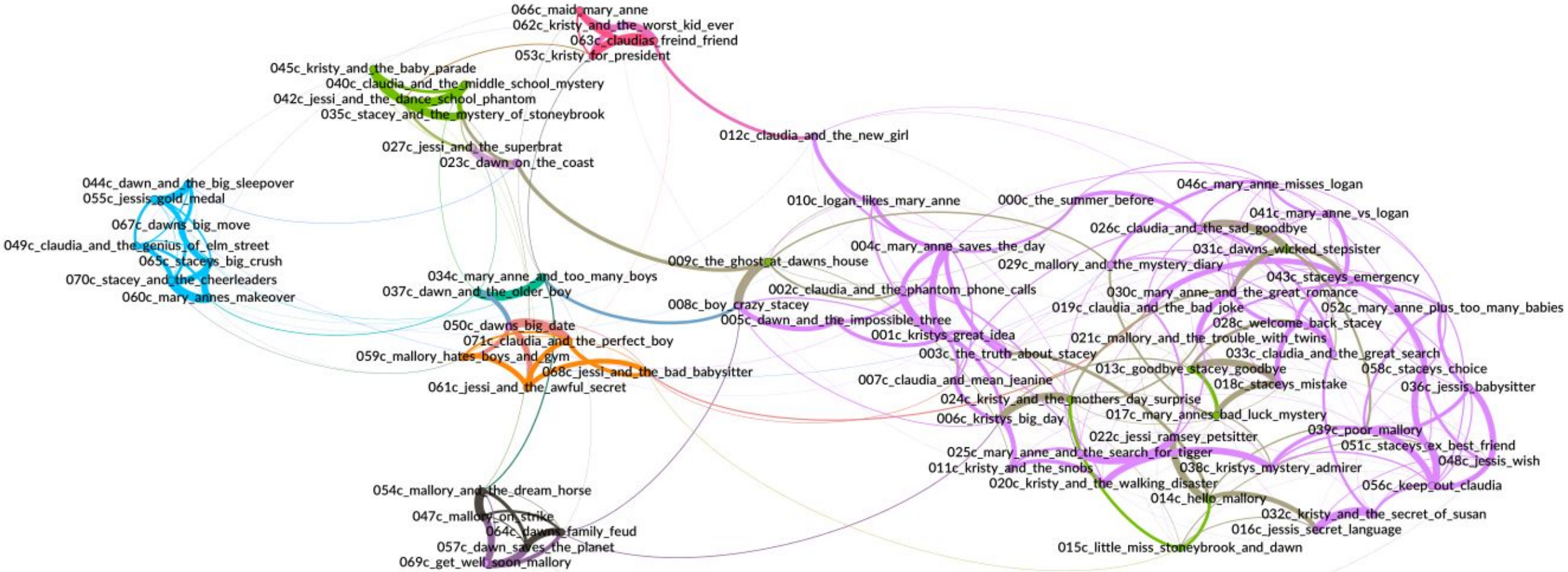
CLSINFRA

Results

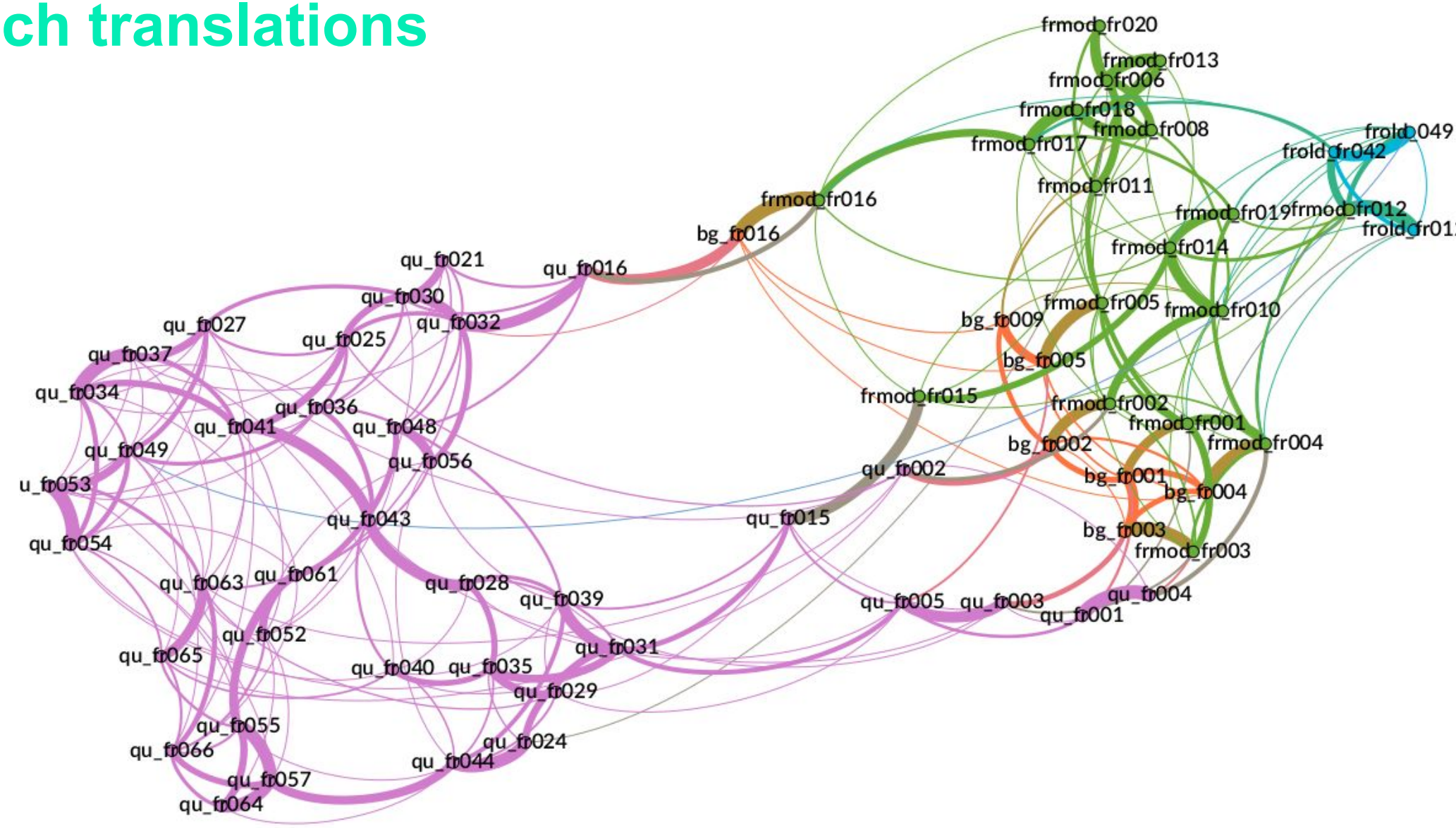
More on:
https://github.com/JoannaBy/BSC-translationese/

Originals (colors indicating (ghost –) writers

CLSINFRA

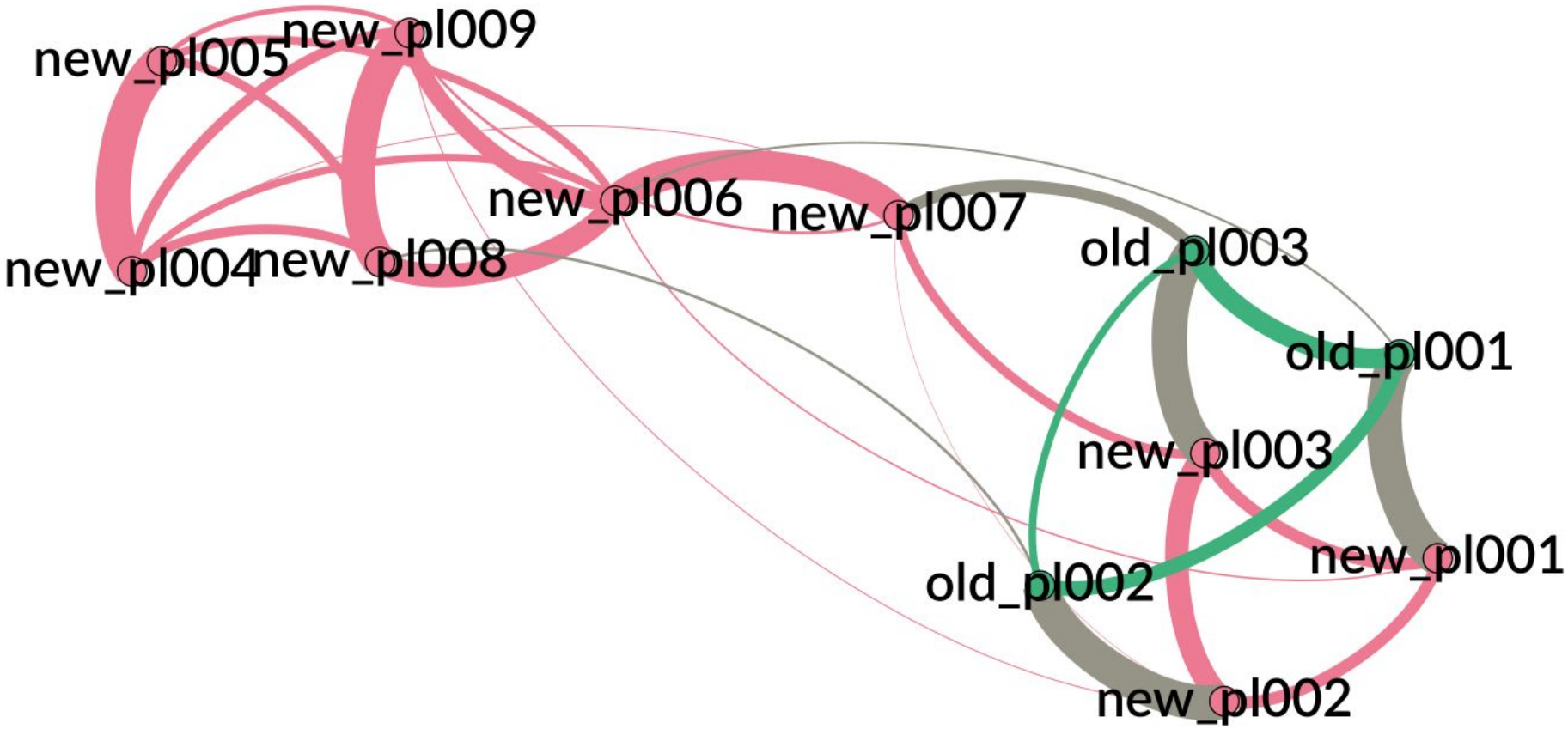


French translations



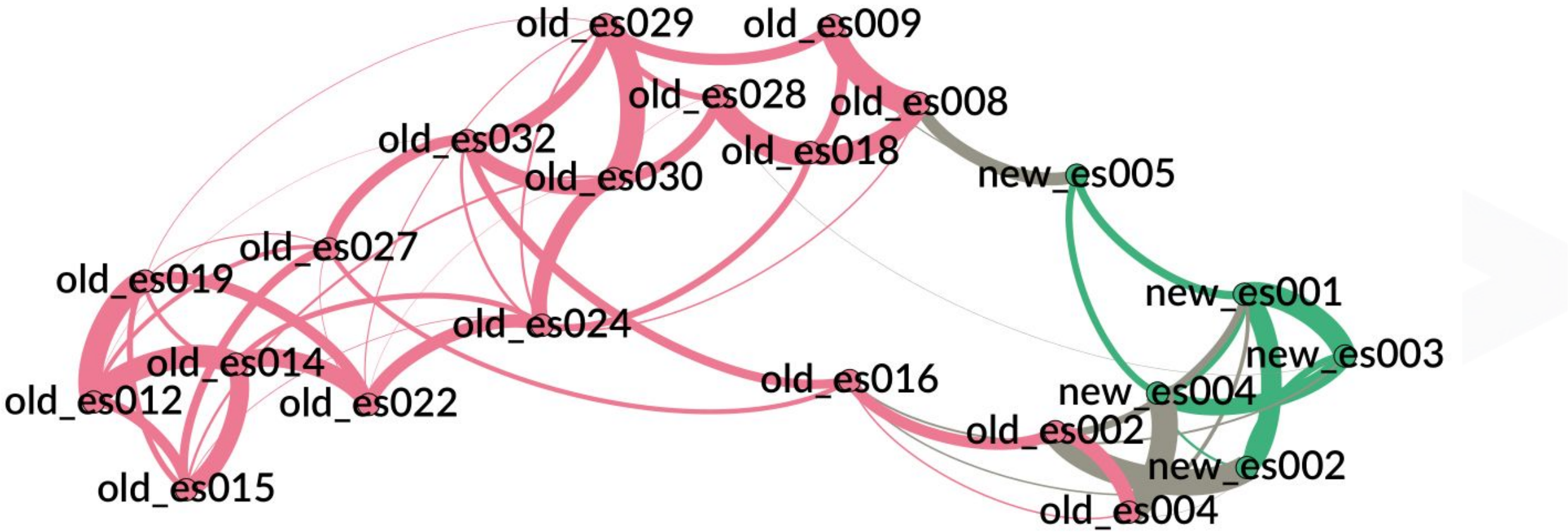
Polish translations

CLSINFRA



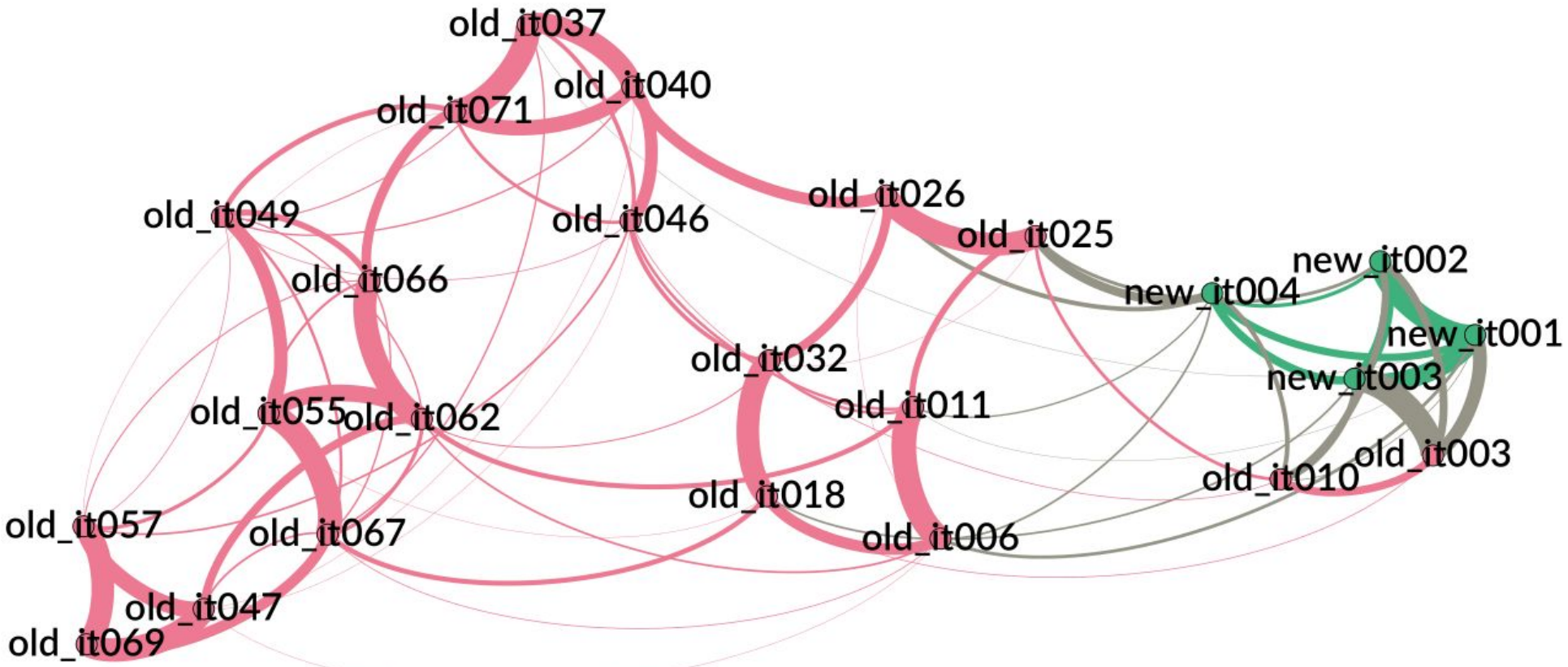
Spanish translations

CLSINFRA



Italian translations

CLSINFRA



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

CLSINFRA COMPUTATIONAL LITERARY STUDIES INFRASTRUCTURE

Rules for designing an experiment

Forming a hypothesis

CLSINFRA

- What is the problem you examine?
- What aspects are to be caught by the study?

Practical part

https://computationalstylistics.github.io/stylo_nutshell/

CLSINFRA

THANK YOU

joanna.byszuk@ijp.pan.pl