Kim naprawdę był Gall Anonim? Zagadnienia statystycznej analizy tekstu

Maciej Eder

Instytut Języka Polskiego PAN

Analiza tekstu

- modelowanie tematyczne (topic modeling)
- analiza emocji (sentiment analysis)
- semantyka dystrybucyjna
- atrybucja autorska
- ...

Pakiety R

- topicmodels, LDA
- tm
- snowballC
- text2vec
- stylo
- ...

Atrybucja autorska (pokrótce)

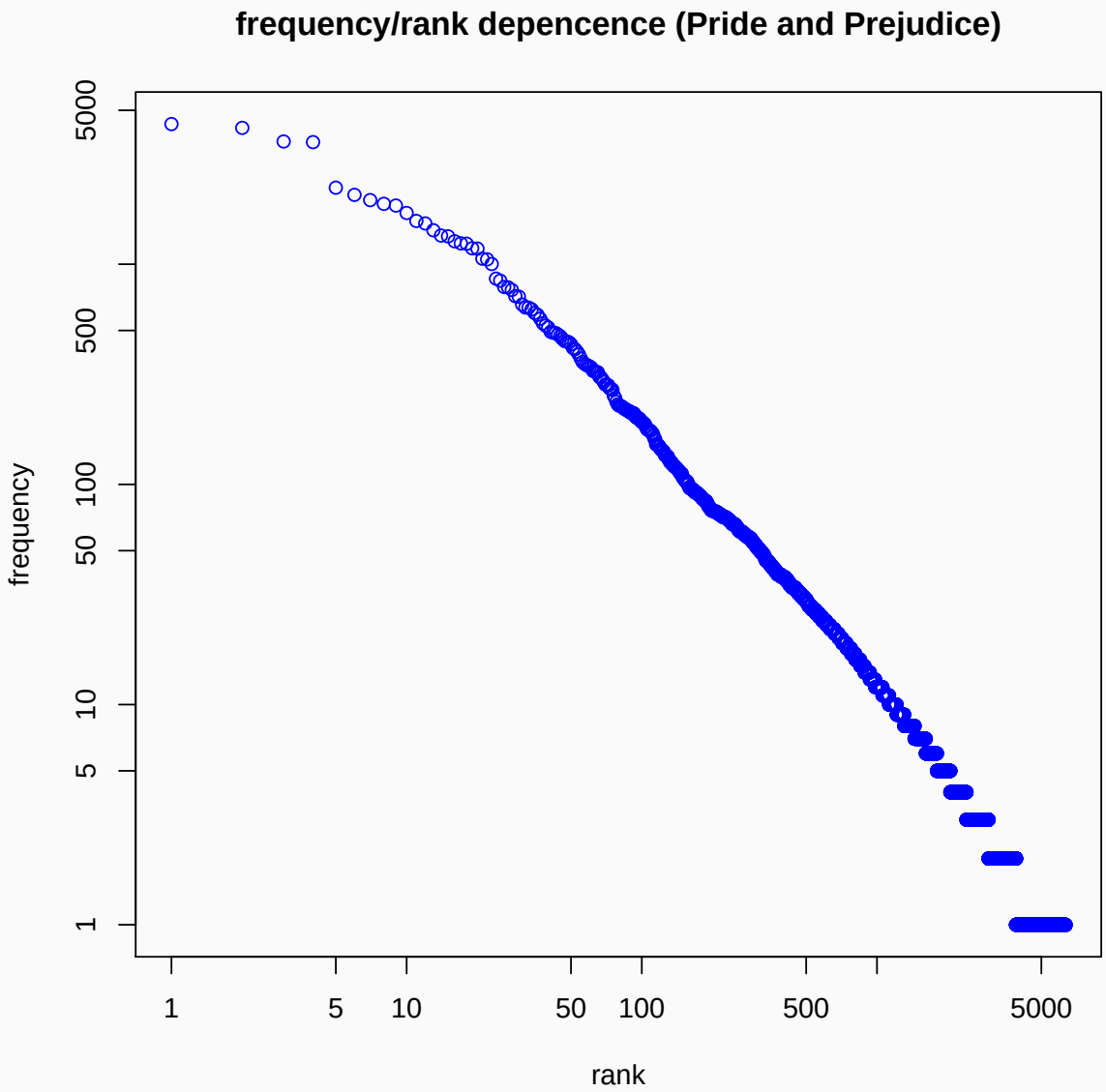
Idea stylistycznego odcisku palca

- nawyki językowe („tiki” stylistyczne)
- cechy języka niedostrzegalne gołym okiem
- poza kontrolą autora (nieświadome)
- odporne na imitację, parodię itp.
- chętnie wybierany znacznik: częstości najczęstszych słów

(1) **Sterne**, *Sentimental*, (2) **Hor.** *Ars*, (3) *Bartachom.*

1.	the	5.7384	et	4.69104	δ’	4.3154
2.	i	3.2138	non	1.26173	καὶ	3.09004
3.	and	3.1292	si	1.22938	ἐς	2.18434
4.	of	3.0518	in	1.13232	δὲ	1.97123
5.	to	2.7640	aut	0.905856	ἦν	1.86468
6.	a	2.4061	qui	0.841152	ὁ	1.17208
7.	it	1.9104	ut	0.776448	οὐ	1.06553
8.	in	1.8838	quid	0.744096	ἐπὶ	1.01225
9.	had	1.1583	nec	0.711744	τὸν	0.745871
10.	was	1.1462	est	0.711744	κατὰ	0.745871
11.	as	1.1100	an	0.420576	τε	0.692595
12.	my	1.0834	ad	0.420576	ἀπ’	0.639318
13.	his	1.0108	quae	0.388224	μὲν	0.639318
14.	he	0.9673	ego	0.388224	ἐπ’	0.586042
...	...	...	...	...	...	...

Prawo Zipfa


frequency/rank depencence (Pride and Prejudice)
frequency
5000
500
100
50
10
5
1
1
5
10
50
100
500
5000
rank

Najczęstsze słowa (= słowa funkcyjne), ...

	ABronte *Agnes*	Austen *Emma*	CBronte *Jane*	Conrad *Lord*	Dickens *Bleak*	...
"the"	3.6747	3.2434	4.1870	5.6837	4.2088	...
"and"	3.9996	3.0403	3.5367	2.5995	3.5770	...
"to"	3.4625	3.2322	2.7532	2.4804	2.8342	...
"I"	3.2254	1.9888	3.8290	2.0441	2.6025	...
"of"	2.3444	2.6750	2.3295	3.4852	2.3944	...
"a"	1.8966	1.9483	2.3872	2.8530	2.1408	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

Niebanalne pytanie: ile właściwie tych słów trzeba policzyć?

Różne rodzaje znaczników stylu

‘Ni vestra auctoritate suffultus, patres pretitulati, vestraque opitulatione fretus fierem, meis viribus in vanum tanti ponderis onus subirem, et cum fragili lembo periculose tantam equoris immensitatem introirem.’

(Anonim tzw. Gall, *Chronica*)

Różne rodzaje znaczników stylu

zwykłe słowa:

‘ni’ ‘vestra’ ‘auctoritate’ ‘suffultus’ ‘patres’ ‘pretitulati’ . . .

bi-gramy słowne:

‘ni-vestra’ ‘vestra-auctoritate’ ‘auctoritate-suffultus’ . . .

bi-gramy literowe:

‘ni’ ‘i-’ ‘-v’ ‘ve’ ‘es’ ‘st’ ‘tr’ ‘ra’ ‘a-’ ‘-a’ ‘au’ ‘uc’ ‘to’ ‘or’ ‘ri’ . . .

tri-gramy literowe:

‘ni-’ ‘i-v’ ‘-ve’ ‘ves’ ‘est’ ‘str’ ‘tra’ ‘ra-’ ‘a-a’ ‘-au’ ‘auc’ ‘uct’ . . .

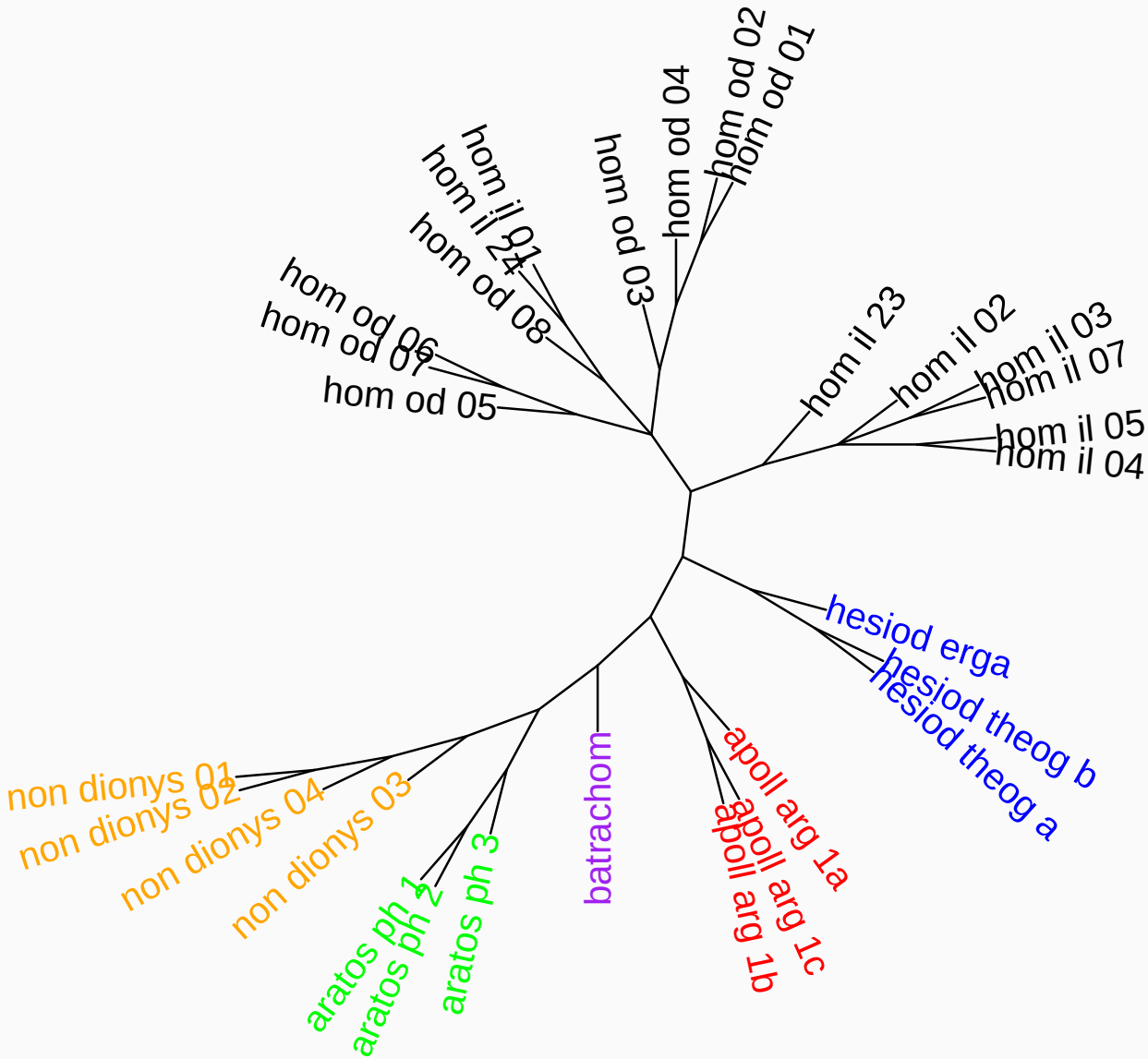
heksa-gramy literowe:

‘ni-ves’ ‘i-vest’ ‘-vestr’ ‘vestra’ ‘estra-’ ‘stra-a’ ‘tra-au’ ‘ra-auct’ . . .

Metody klasyfikacji nienadzorowane/nadzorowane

- Principal Components Analysis
- Multidimensional Scaling
- Bootstrap Consensus Networks
- k-Nearest Neighbors
- Support Vector Machines
- Nearest Shrunken Centroids
- ...

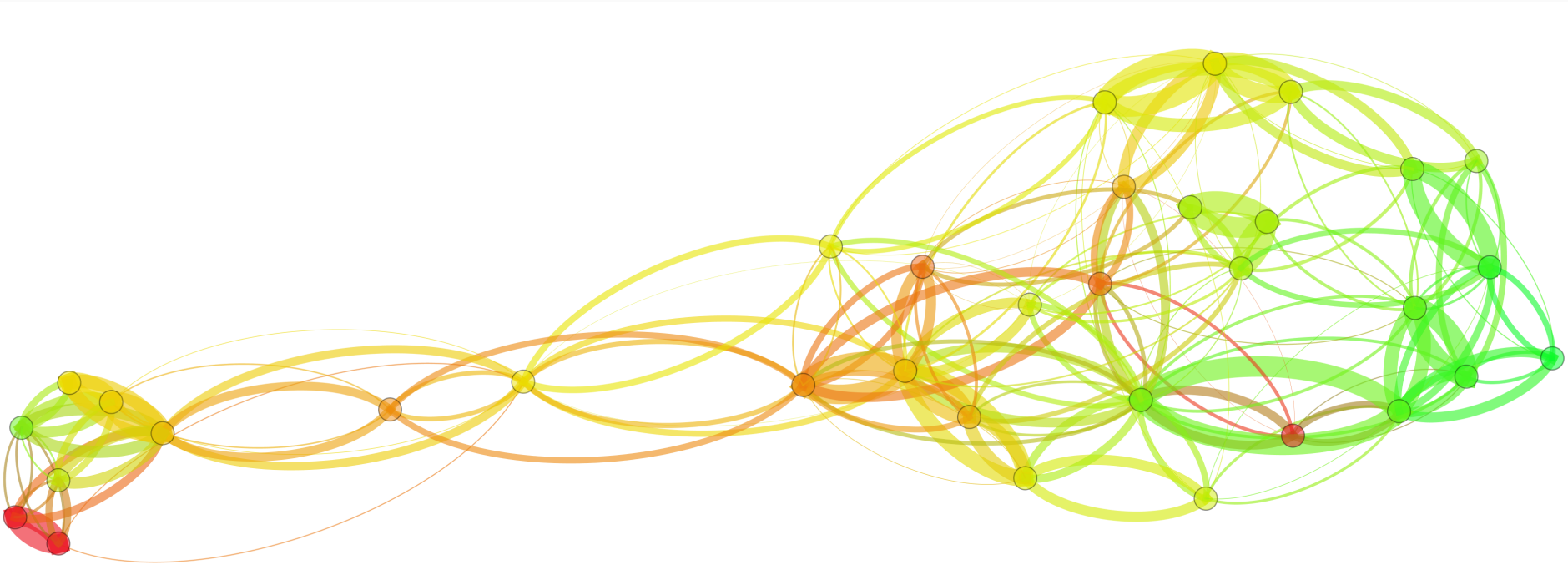
Przykład: kto napisał *Batrachomyomachię*?



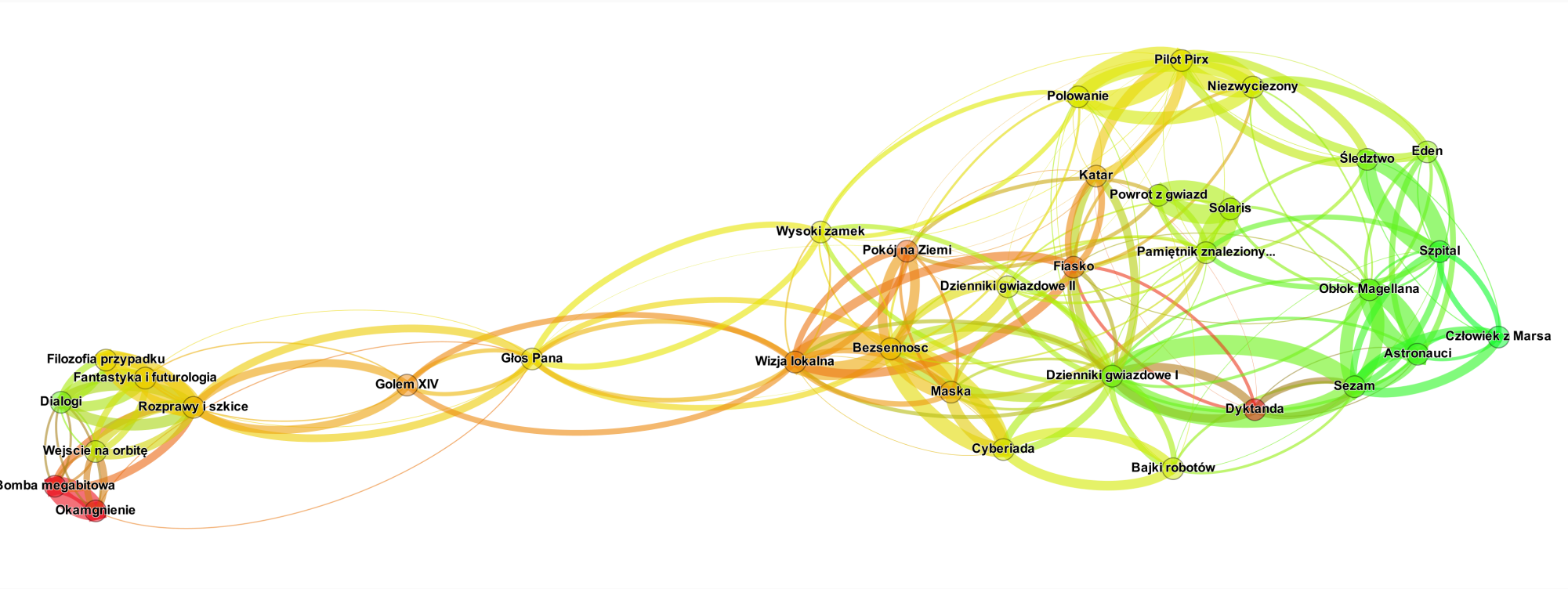
Czy tylko sygnał autorski?

Tekst jako fenomen wielowarstwowy

Stanisław Lem: sygnał gatunku oraz chronologia



Stanisław Lem: sygnał gatunku oraz chronologia

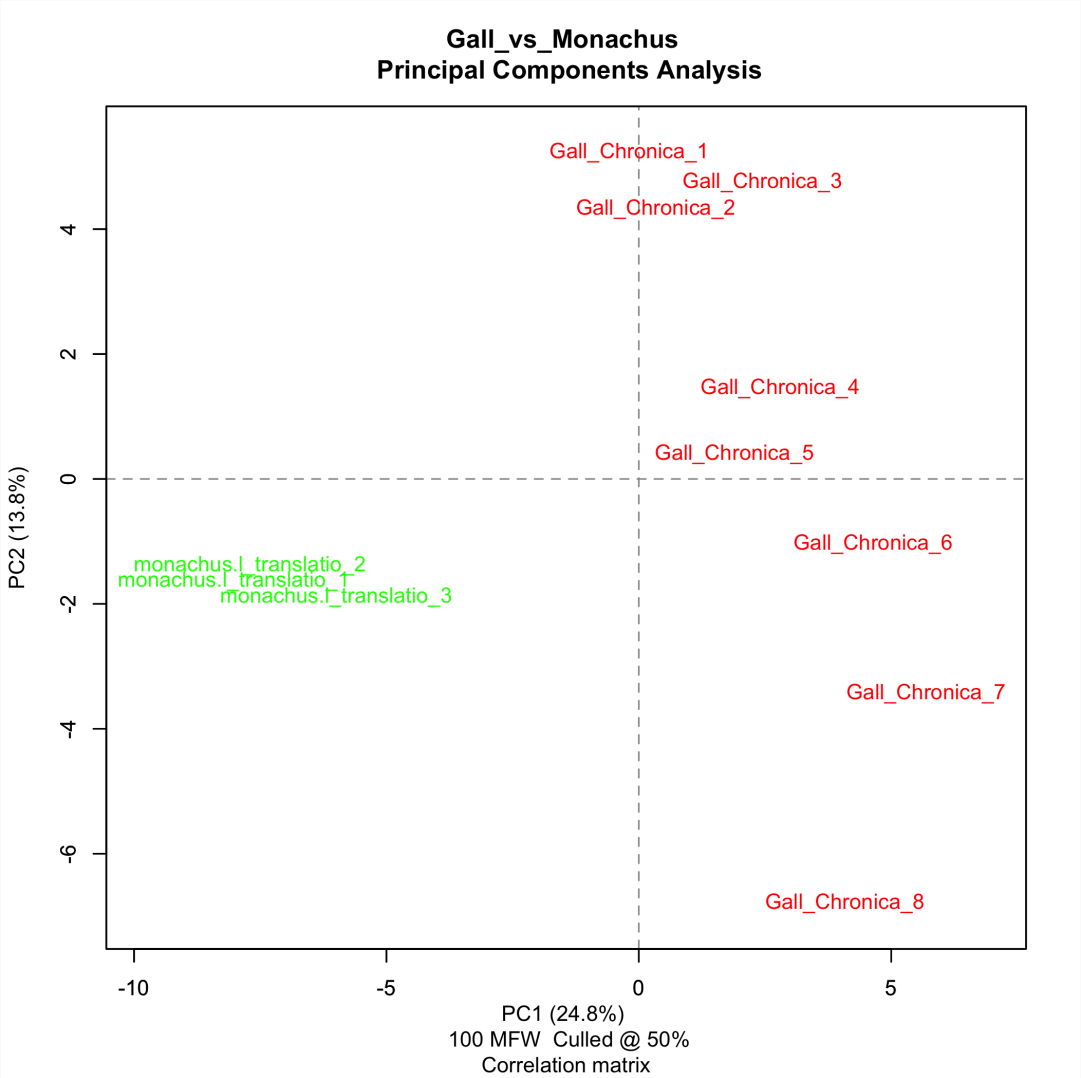


Gallus Anonymus czy Italus Anonymus?

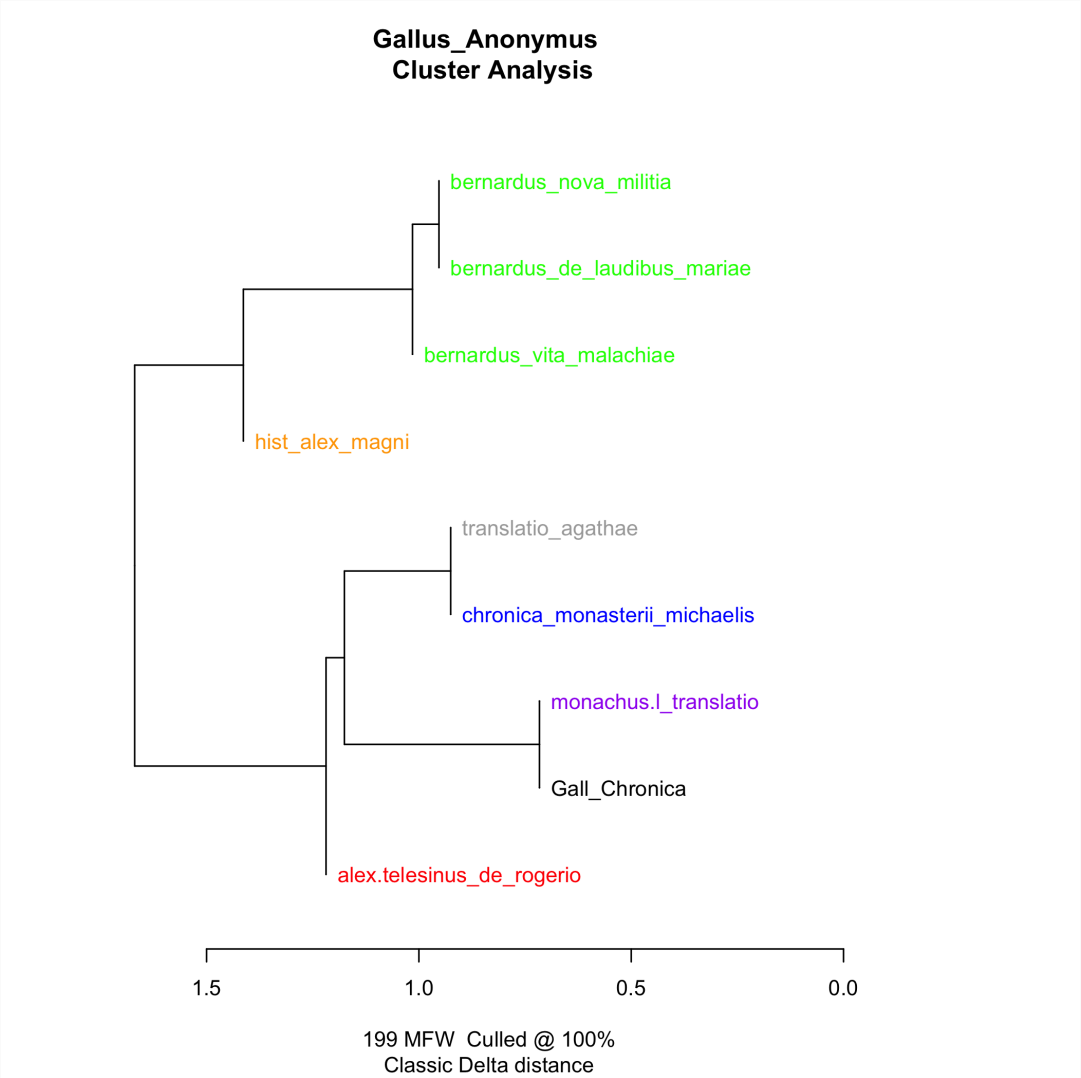
Kronika polska (Chronica Polonorum)

- nazwiska autora nie znamy
- wiadomo, że obcokrajowiec („exul apud vos et peregrinus”)
- hipoteza francuska („Gall” Anonim): Marcin Kromer w XVI w.
- hipoteza węgierska: Bruckner
- hipoteza wenecka: Jasiński

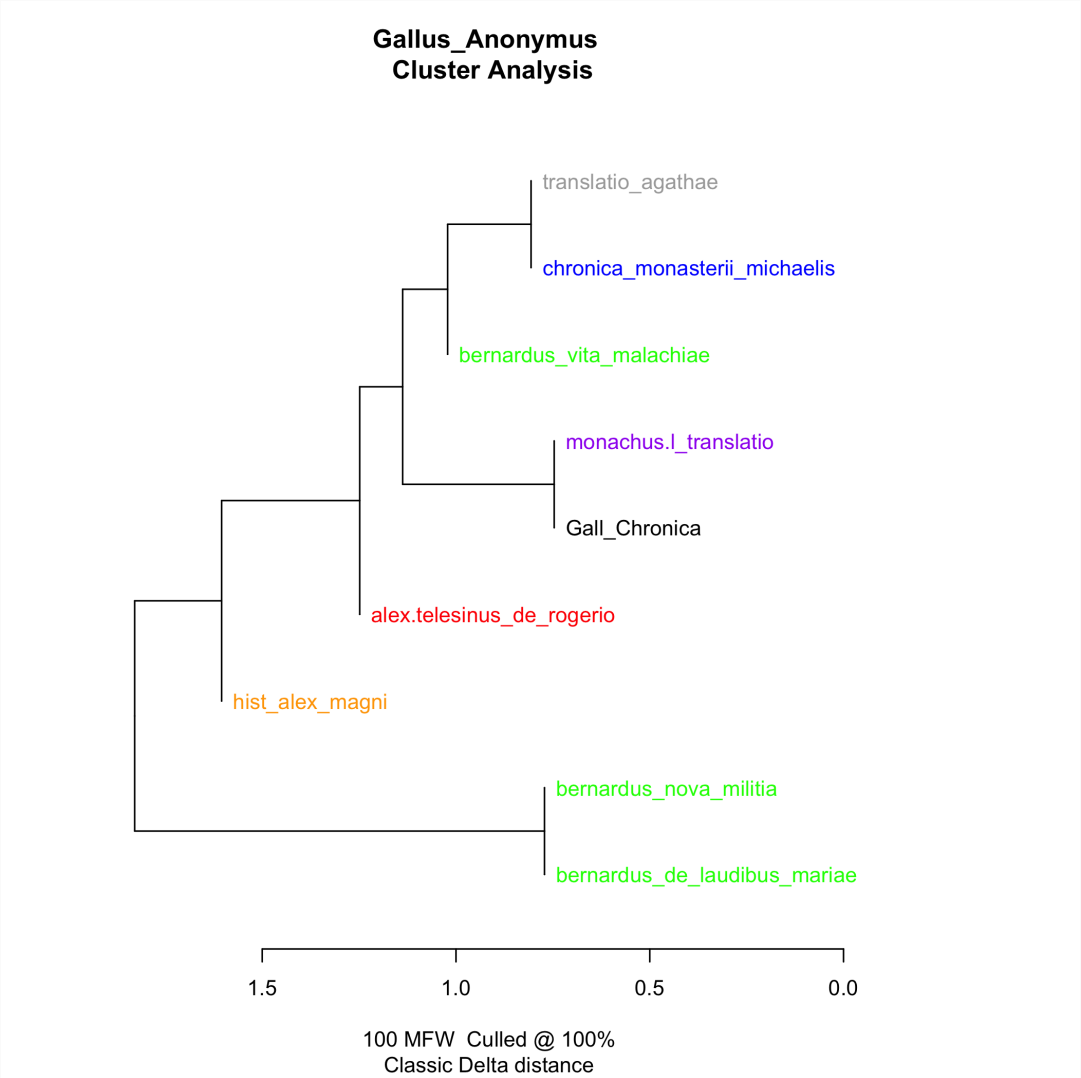
Gallus vs. Italus: pierwsze porównanie



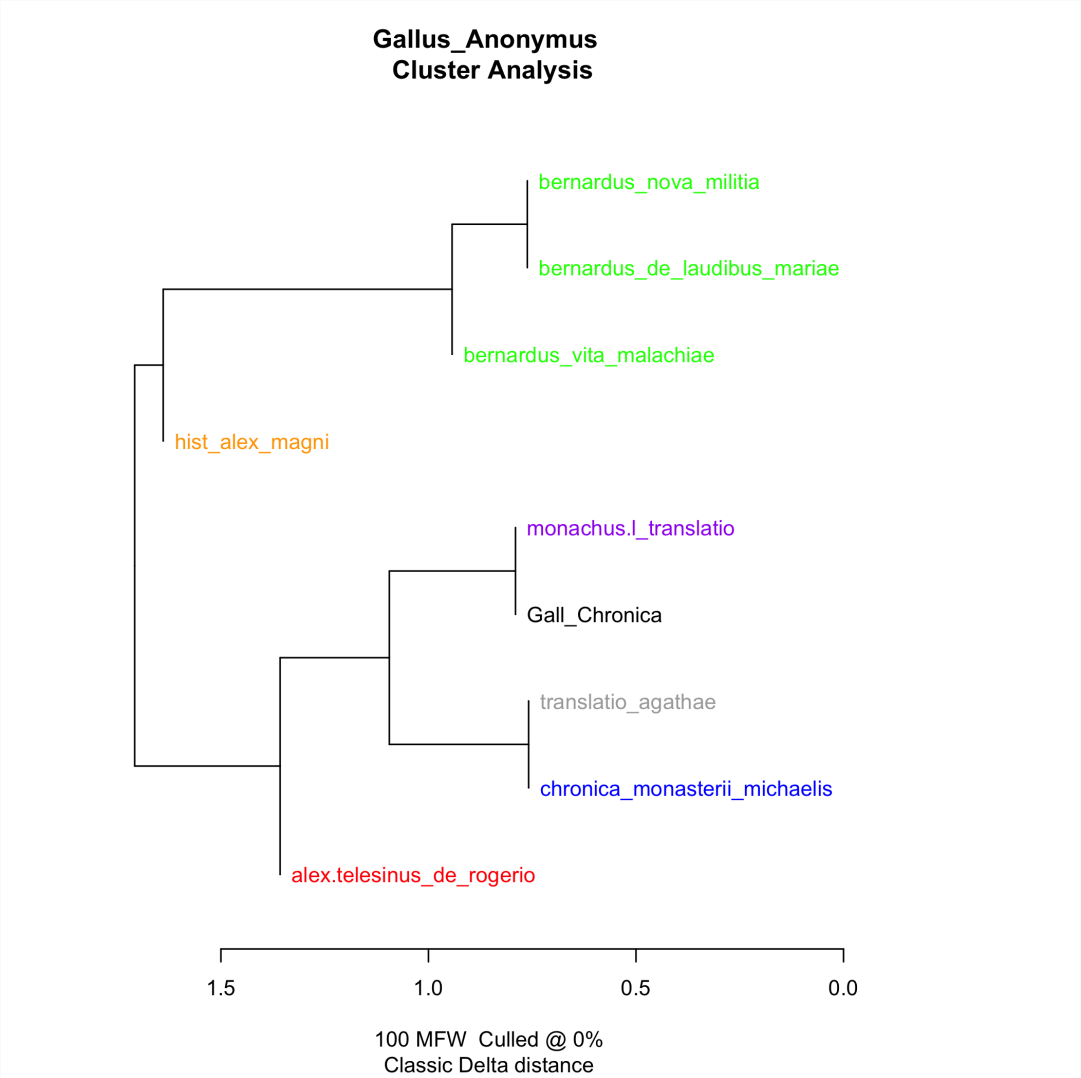
Gall na tle kilku tekstów z epoki



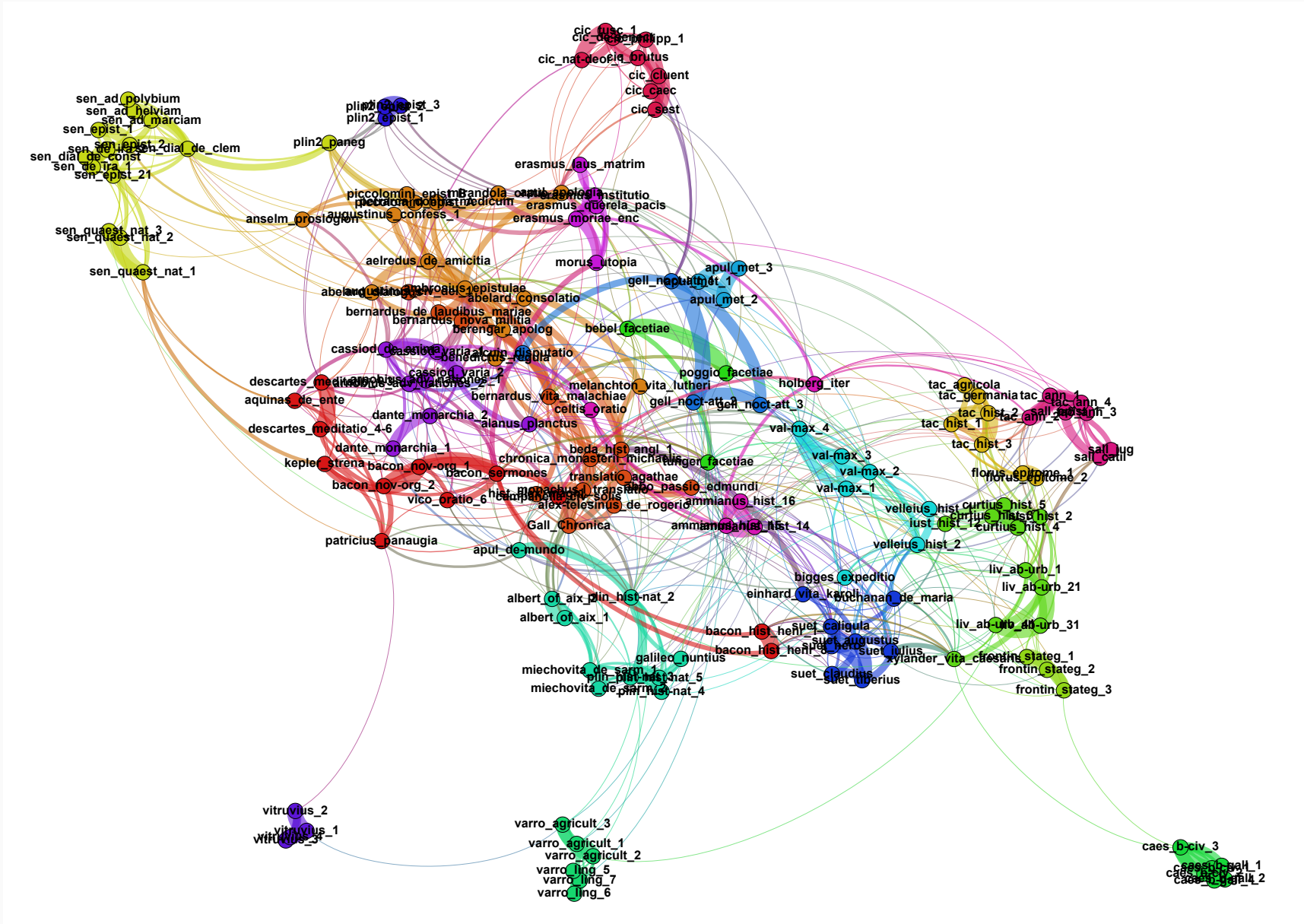
Gall konsekwentnie obok Mnicha z Lido



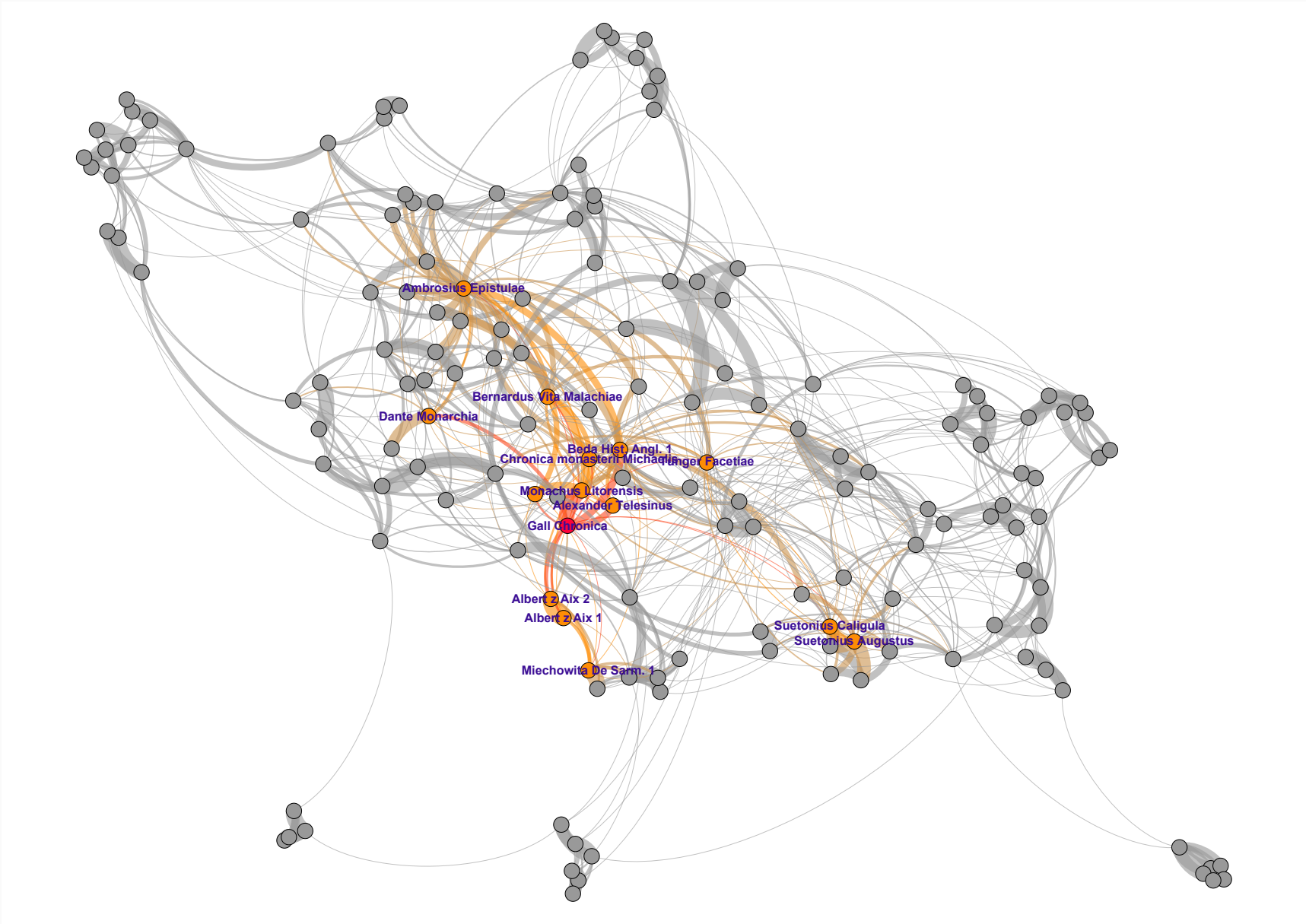
Gall konsekwentnie obok Mnicha z Lido



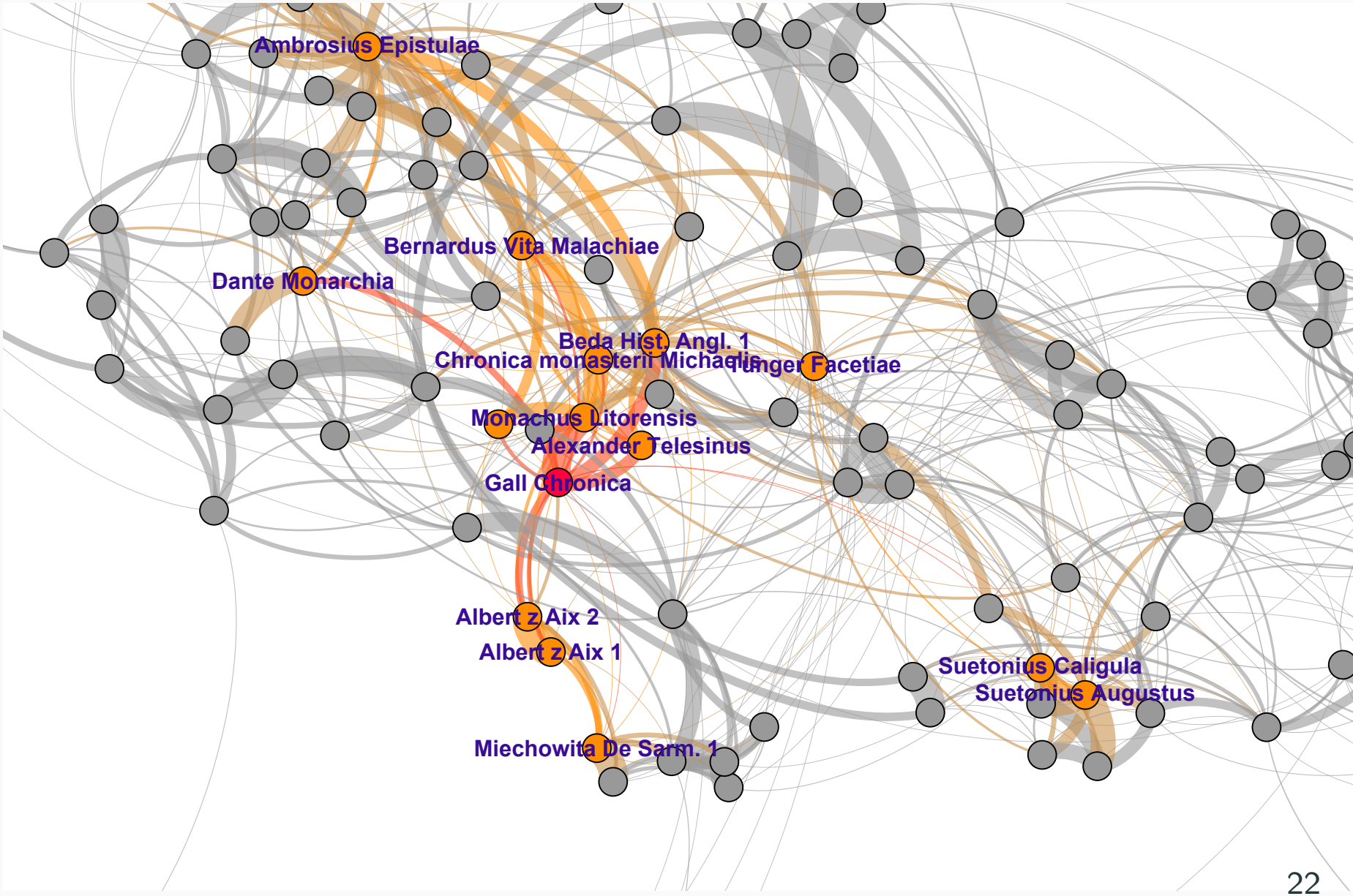
Gall na tle 160 tekstów łacińskich



Gall konsekwentnie obok Mnicha z Lido

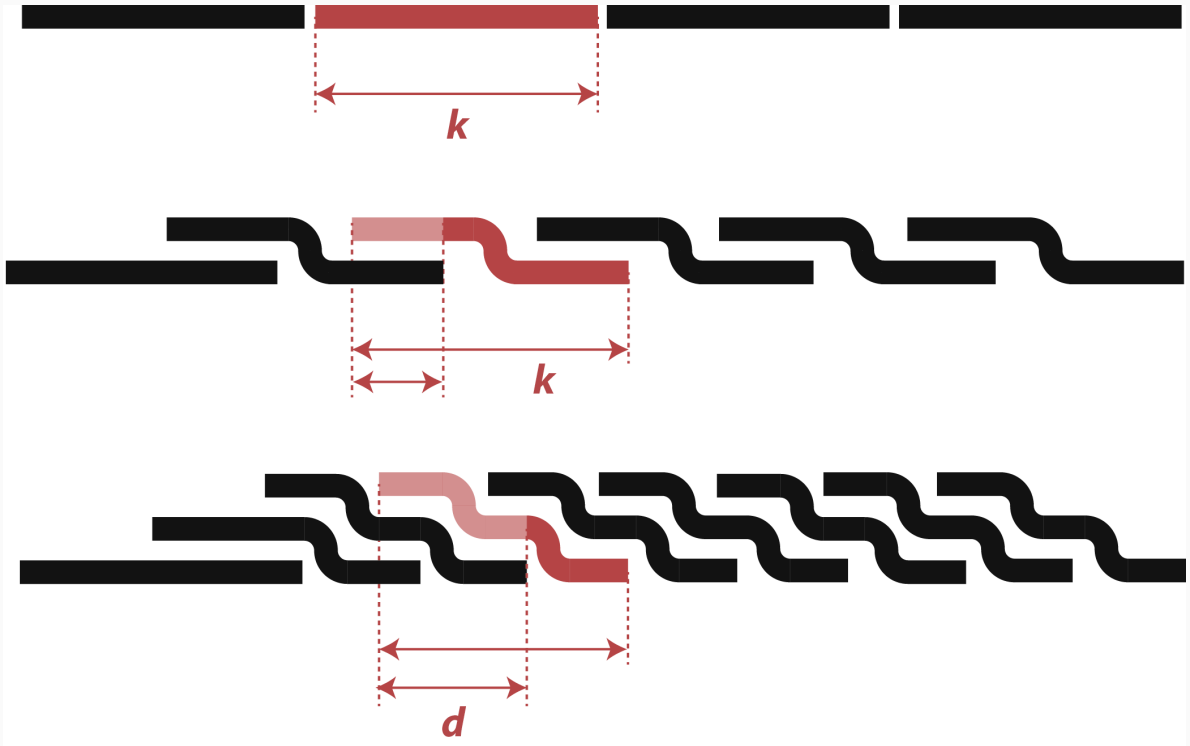


Gall konsekwentnie obok Mnicha z Lido



Analiza sekwencyjna a stylometria

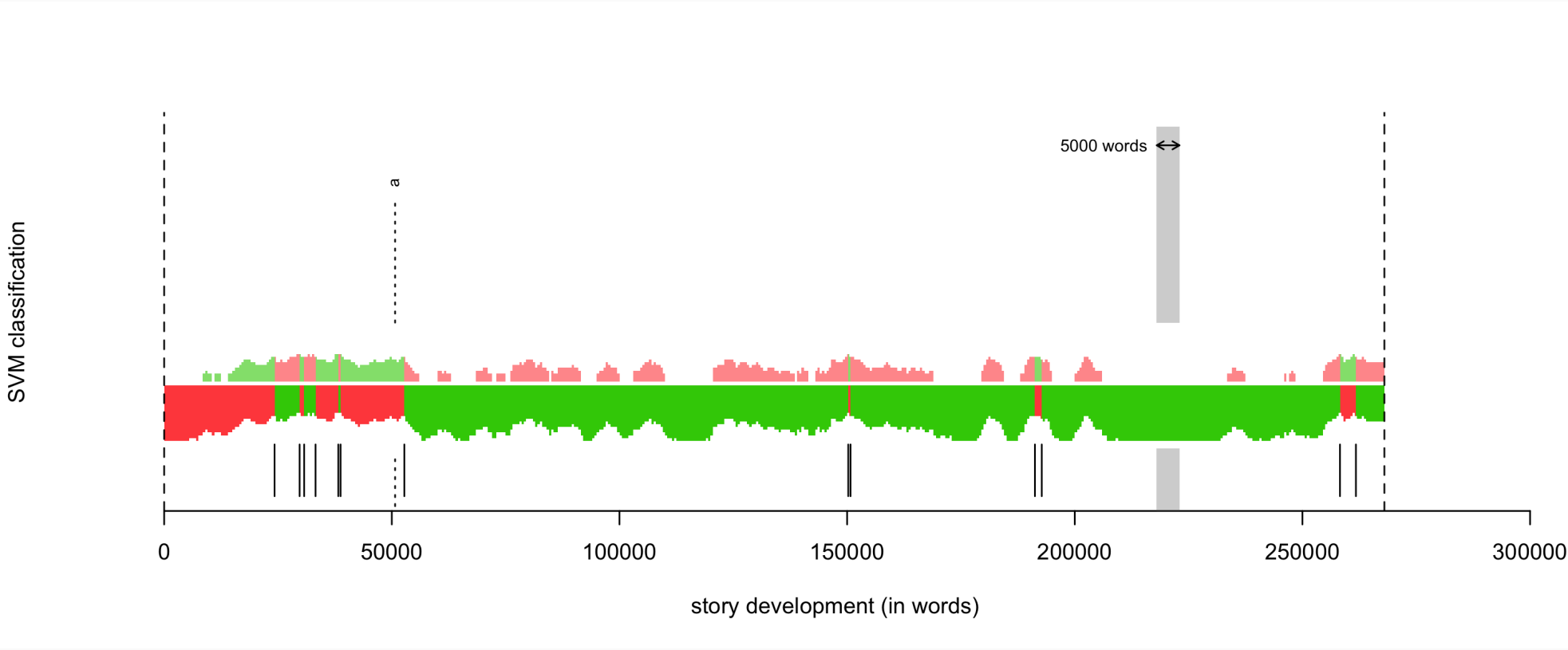
Analiza sekwencyjna: próbkowanie



Romans o Róży (*Roman de la Rose*)

- XIII-wieczny starofrancuski poemat alegoryczny...
- ... napisany przez dwóch autorów:
  - Wilhelm z Lorris (ca. 1230)
  - Jan z Meun (ca. 1275)
- Wiemy, gdzie nastąpiła zmiana (w. 4058).

Romans o Róży i klasyfikacja nadzorowana (SVM)



Harper Lee i Truman Capote

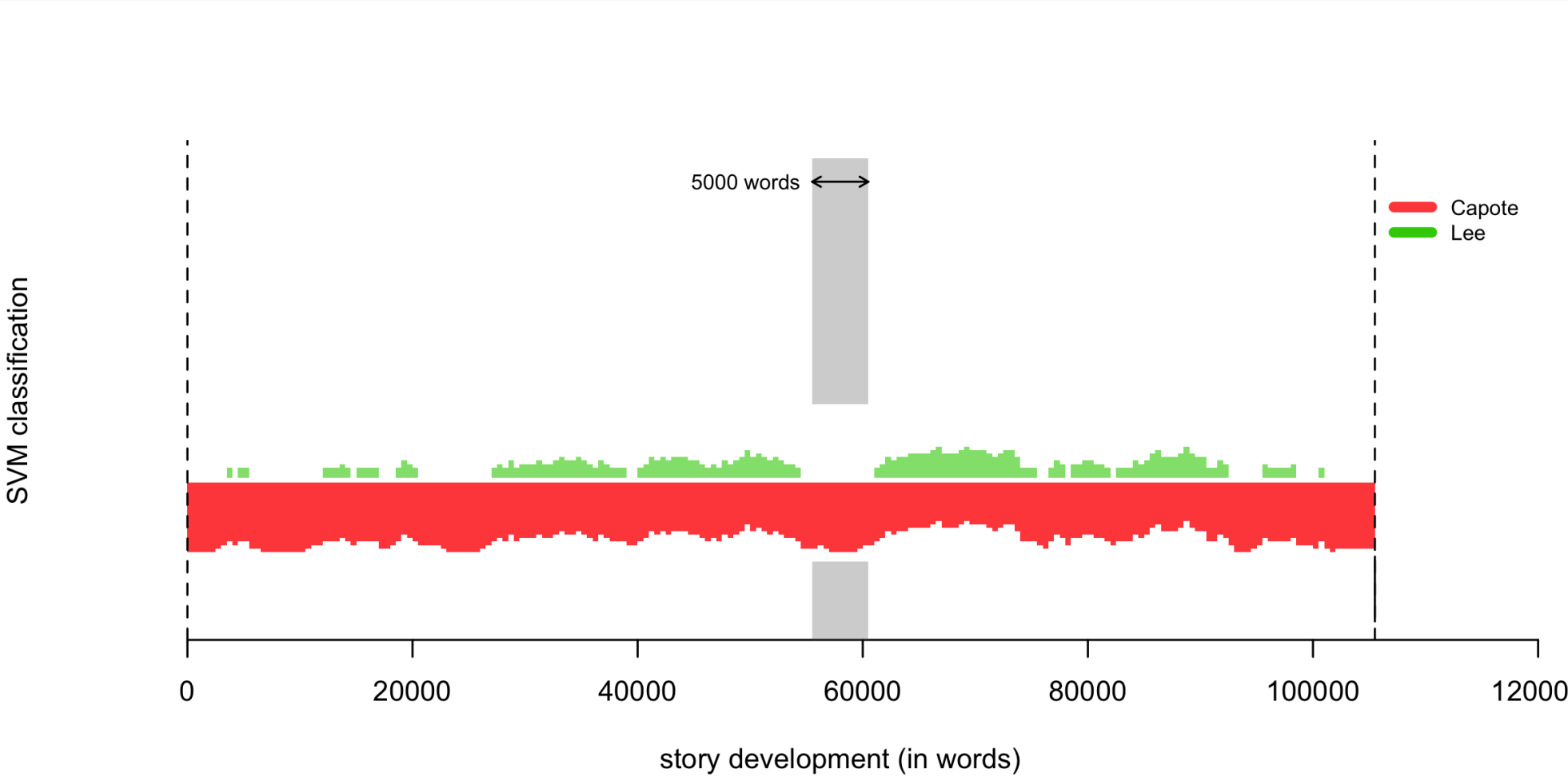
Zabić drozda vs. *Go Set a Watchman*

- 14 lipca 2015 ukazała się powieść *Go Set a Watchman*:
  - druga powieść Harper Lee, opublikowana po 50 latach
  - *Zabić drozda* (1960): nagroda Pulitzera
  - mocno przeredagowana, pod silnym wpływem Tay Hohoff
- Zarzuty, że prawdziwym autorem *Drozda* Truman Capote
- Zarzuty, że Truman Capote nie napisał sam *Z zimną krwią*

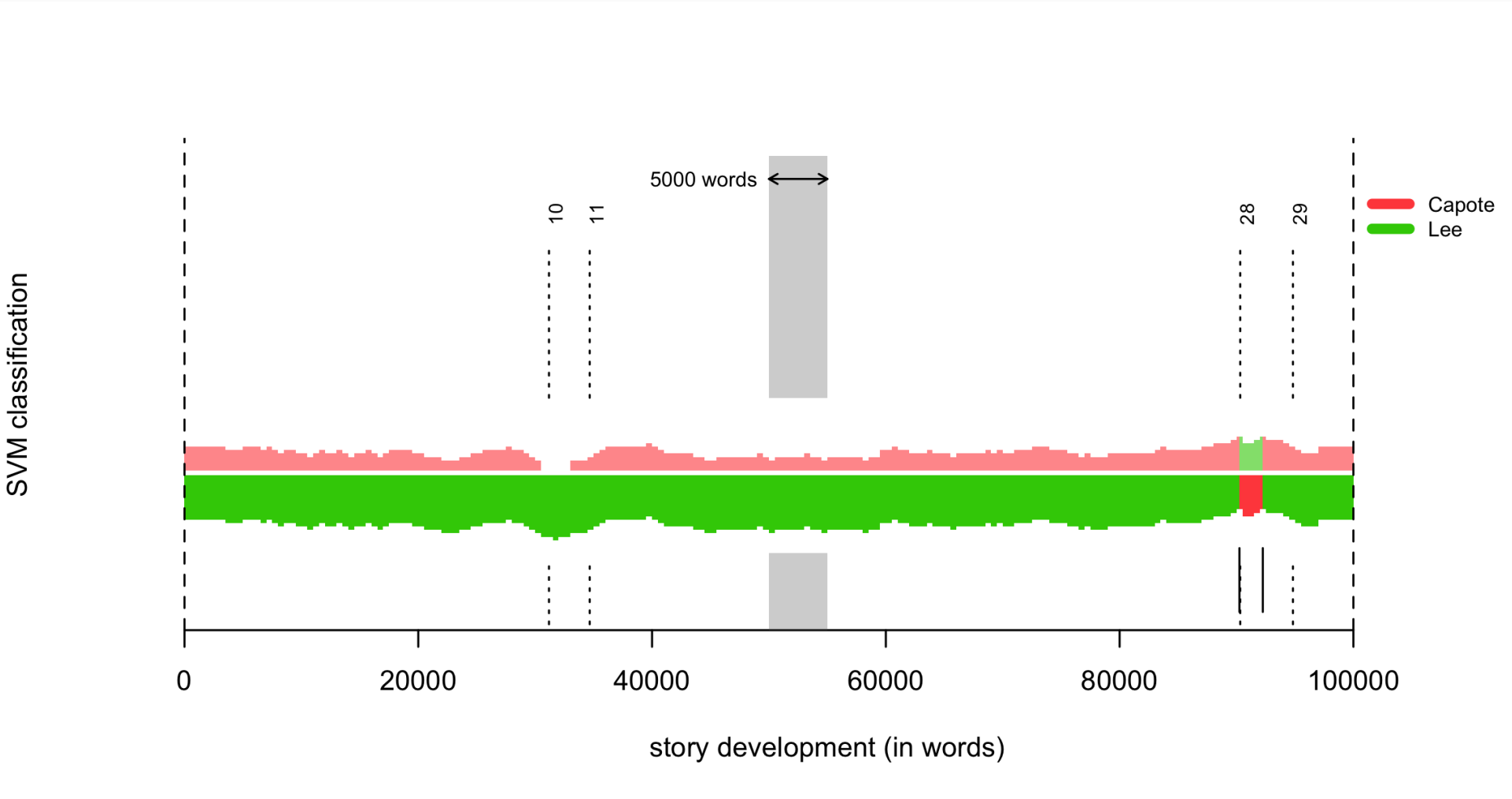
Grupa krakowska o Harper Lee

- Wall Street Journal (17 lipca 2015)
- Gazeta Wyborcza (31 lipca 2015)
- New York Review of Books (24 września 2015)
- Corriere della Sera (17 lutego 2016)

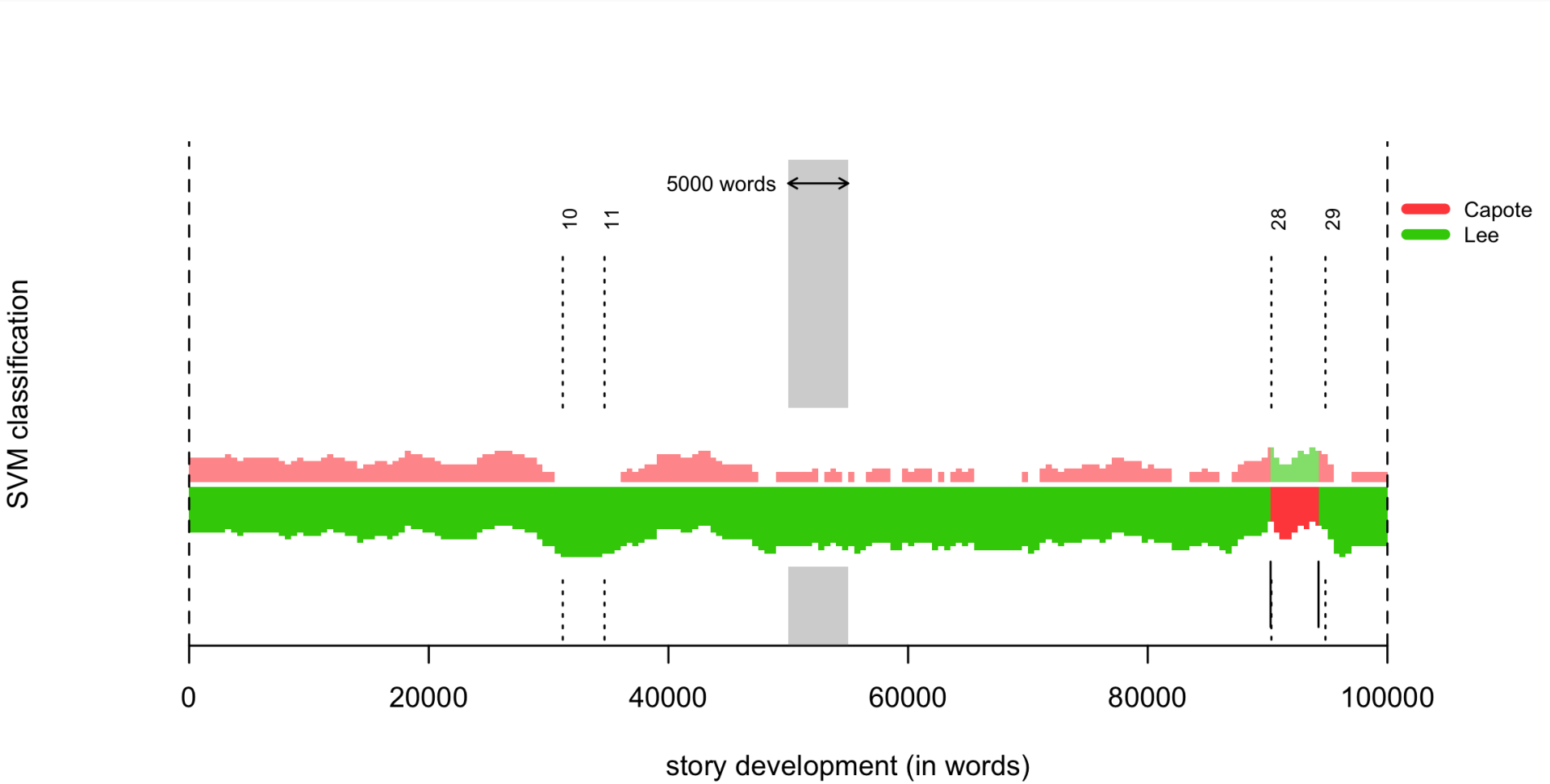
Z zimną krwią pokrojone na segmenty (100 MFW)



Zabić drozda pokrojone na segmenty (300 MFW)



Zabić drozda pokrojone na segmenty (100 MFW)



WhyR?

Analiza tekstu w R: pakiet "stylo"

sites.google.com

10 Computational 01
01 Stylistics 0101000
11 Group 011010110

Search this site

main page

Navigation

- main page
  - more photos
- corpora
- papers and articles
  - preprints
- projects
  - Computer Methods in Textual Studies 2014
  - Go Set A Watchman while we Kill the Mockingbird In Cold Blood
  - testing big dendrograms
  - Testing consensus networks
  - testing rolling delta
  - Testing rolling stylometry
  - translationese
- stylo R package
  - installation hints
  - scripts
  - scripts: obsolete

Stylometry@Kraków

Computational Stylistics Group is a cross-institutional research team focused on computer-assisted text analysis, stylometry, authorship attribution, sentiment analysis, and the like stuff. The group is based in Krakow and Antwerp, at the Institute of Polish Language (Polish Academy of Sciences), the Jagiellonian University, and the University of Antwerp. CSG is a member of the Federation of Stylometry Labs (FoSL).

latest news: the version **0.6.4** of the R package "stylo" released! Click here for further details.

This HOWTO of the package "stylo" lets you make your first stylometric analysis in no time. Also, it might be worthwhile to visit this discussion group.

If you find the package "stylo" useful and plan to publish your results, please consider citing the following paper:
Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. *R Journal*, **8**(1): 107–121, url: https://journal.r-project.org/archive/2016-1/eder-rybicki-kestemont.pdf

Maciej Eder (center) is Director of the Institute of Polish Language at the Polish Academy of Sciences, and Associate Professor at the Institute of Polish Studies at the Pedagogical University of Kraków, Poland. He is interested in European literature of the Renaissance and the Baroque, classical heritage in early modern literature, and scholarly editing (his most recent book is a critical edition of 16th-century Polish translations of *Dialogue of Salomon and Marcolf*). A couple of years ago while doing research on anonymous ancient texts, Eder discovered the fascinating world of computer-based stylometry and non-traditional authorship attribution. His work is now focused on a thorough re-examination of current attribution methods and applying them to non-English languages, e.g. Latin and Ancient Greek.

Jan Rybicki (left) is Assistant Professor at the Institute of English Studies,