Computational Literary Text Analysis (1): an Overview

Erasmus+ Lectures

Kraków, March/April 2016

Christof Schöch
CLiGS - University of Würzburg

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

CLiGS

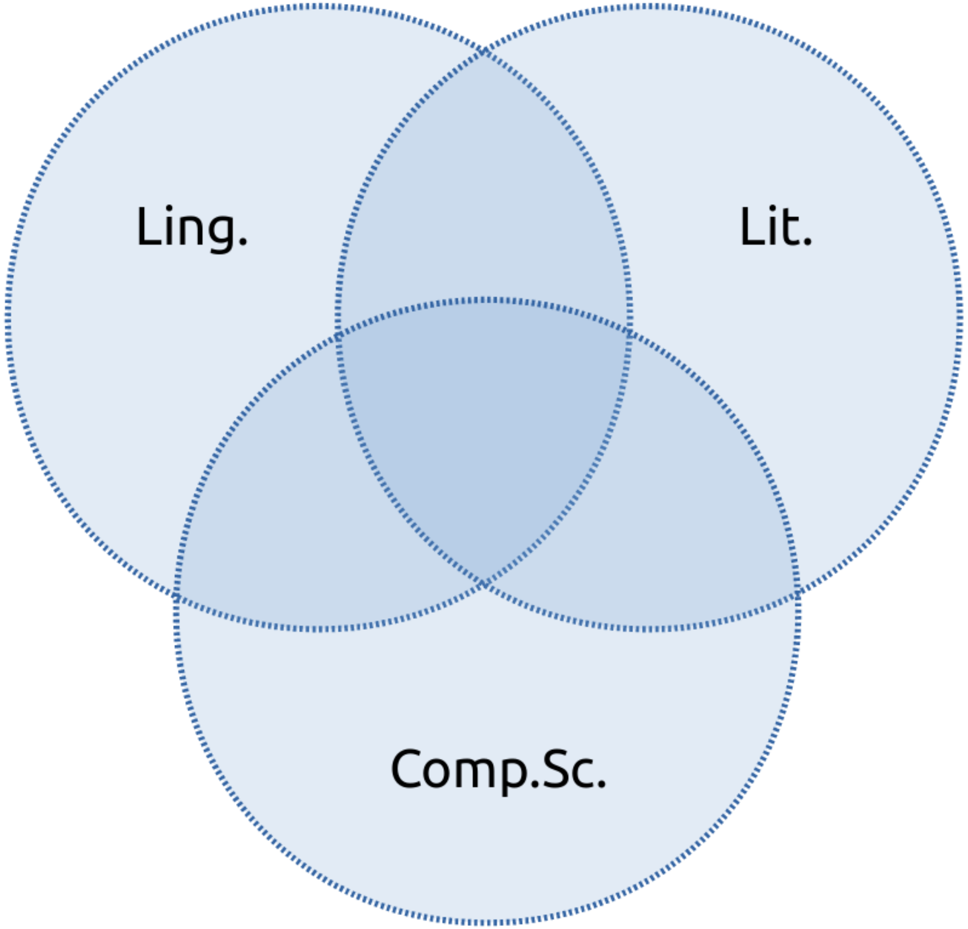
GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Overview

1. What is computational literary text analysis?
2. Stylometric Similarity Analysis
3. Machine Learning in Computational Narratology
4. Thematic Analysis with Topic Modeling
5. Conclusion

1. What is computational literary text analysis?

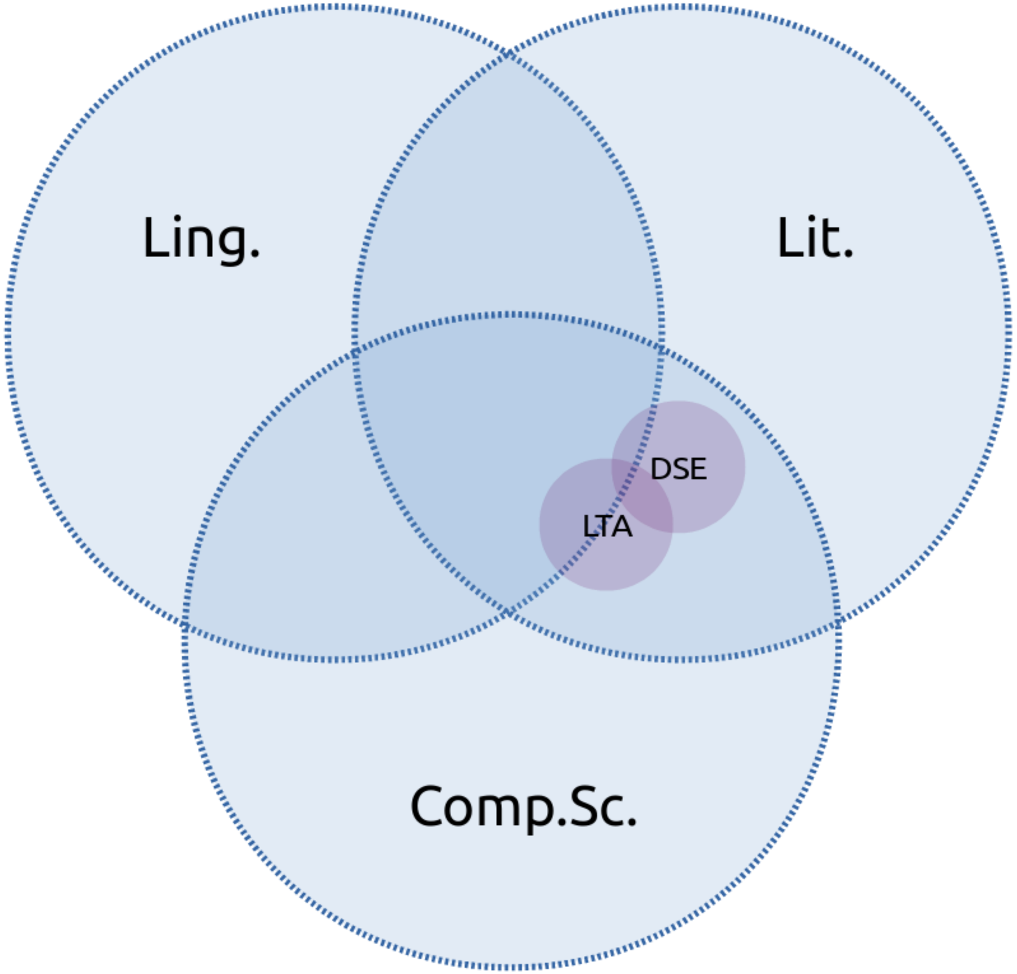
Literary Computing



Literary Computing

- Creating Texts: Digitisation, Digital Scholarly Editing (TEI), Text Collections
- Analysing Texts: Style, Content, Space, Time, Characters; Authors, Genres, Periods, etc.

Literary Text Analysis



(DSE = Digital Scholarly Editing, LTA = Computational Literary Text Analysis)

Some distinctions

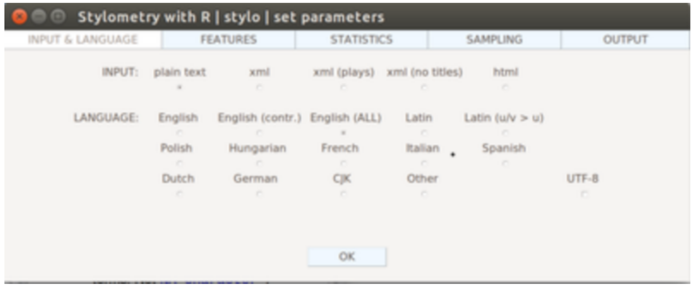
Some areas of LTA

Some techniques

2. Stylometric Similarity Analysis

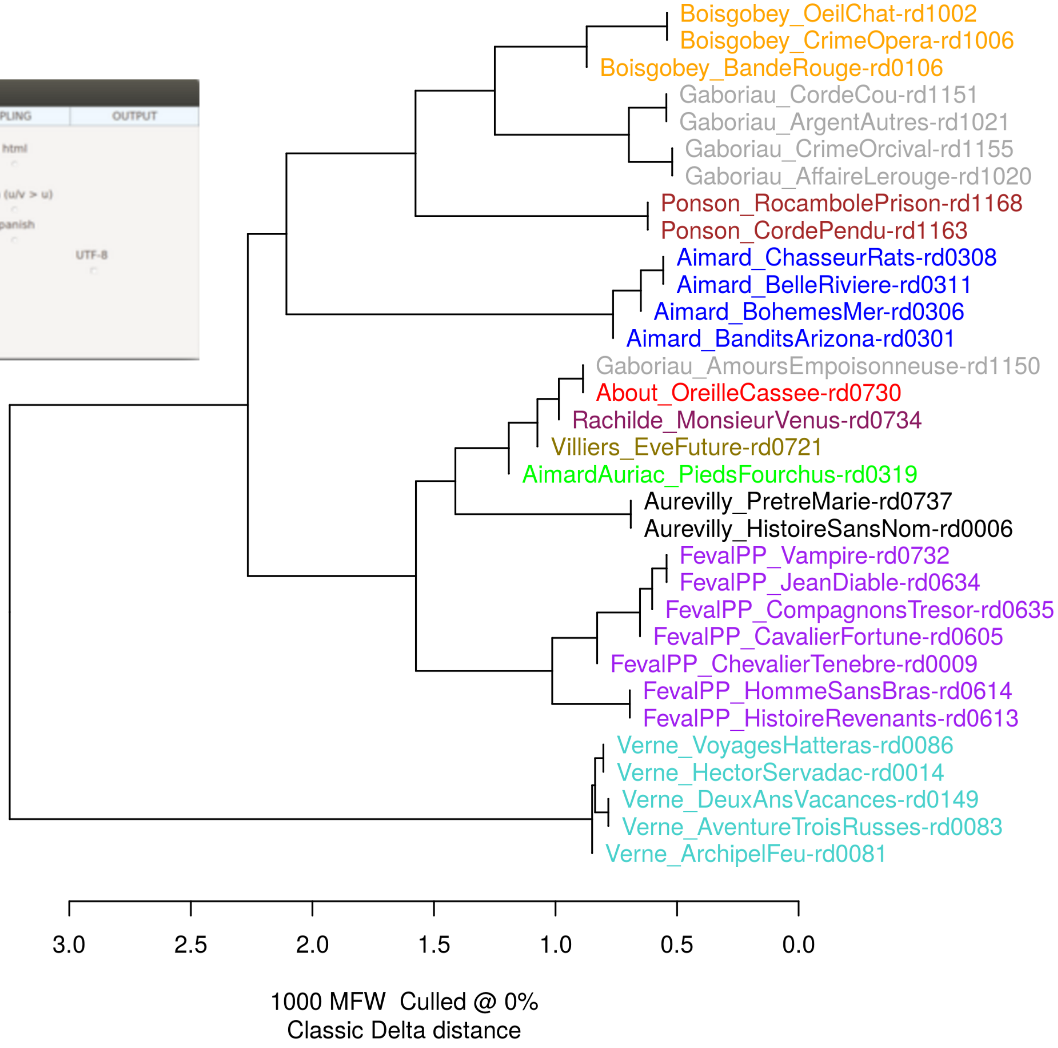
(With thanks to Maciej Eder, Jan Rybicki, Stefan Evert, Fotis Jannidis, Thorsten Vitt, Steffen Pielström, ...)

stylo for R



INPUT: ● plain text ○ xml ○ xml (plays) ○ xml (no titles) ○ html

LANGUAGE: ○ English ○ English (contr.) ● English (ALL) ○ Latin ○ Latin (u/v > u)
○ Polish ○ Hungarian ○ French ○ Italian ○ Spanish
○ Dutch ○ German ○ CJK ○ Other ○ UTF-8

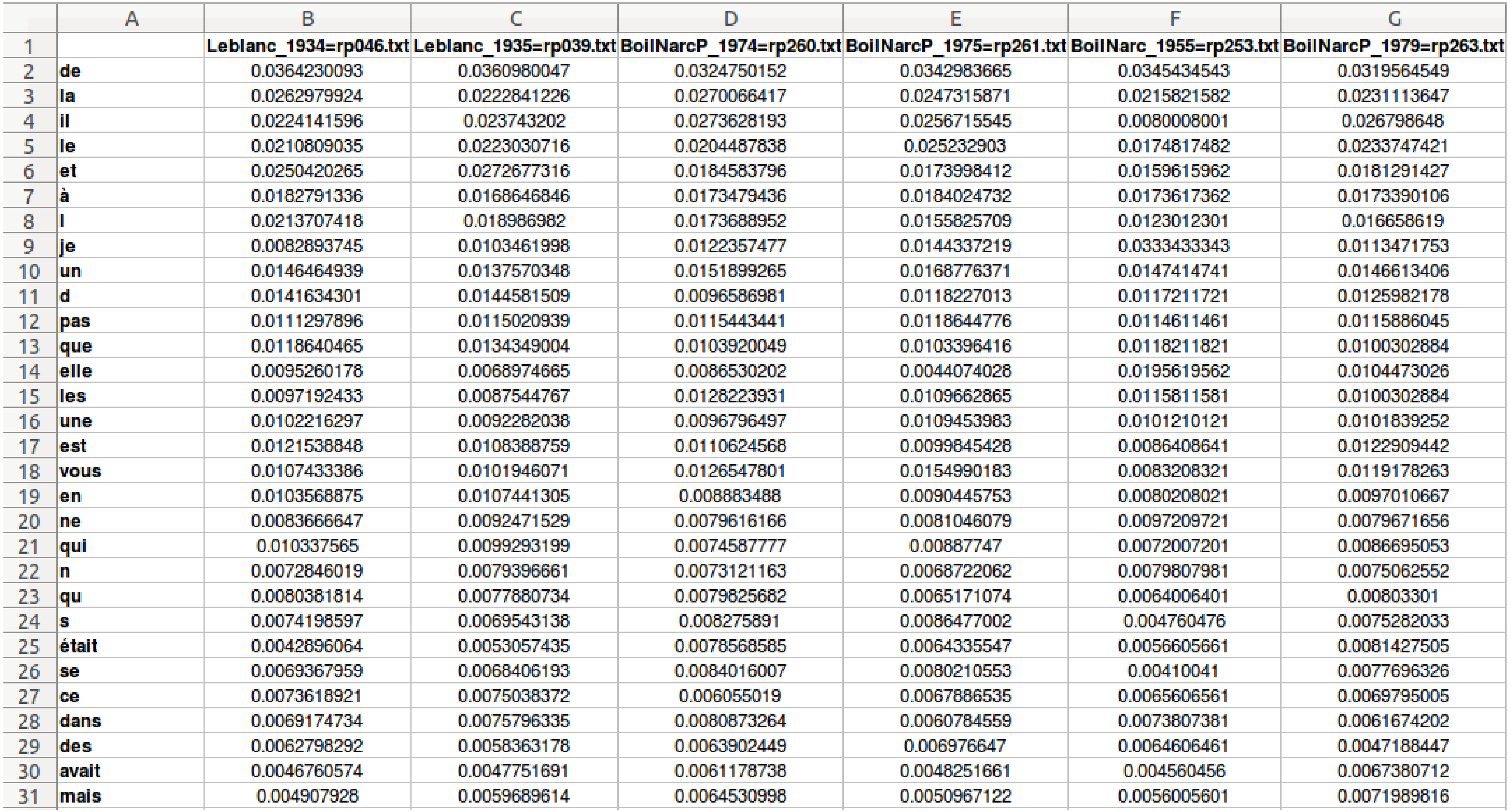


Stylometric Similarity Analysis

- Clearly quantitative method
- Usually using surface features
- Can be used for clustering or classification
- There is also a GUI ;-)

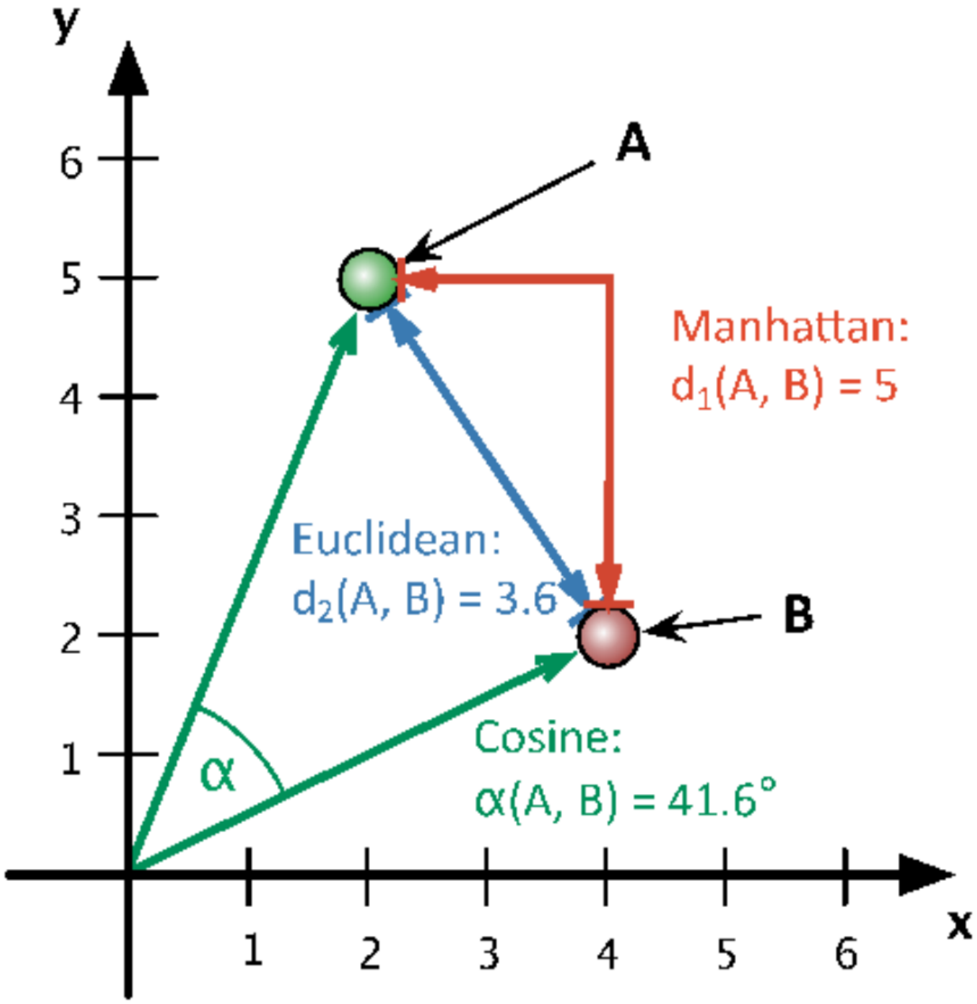
How does stylometric similarity work?

Texts as word frequency vectors



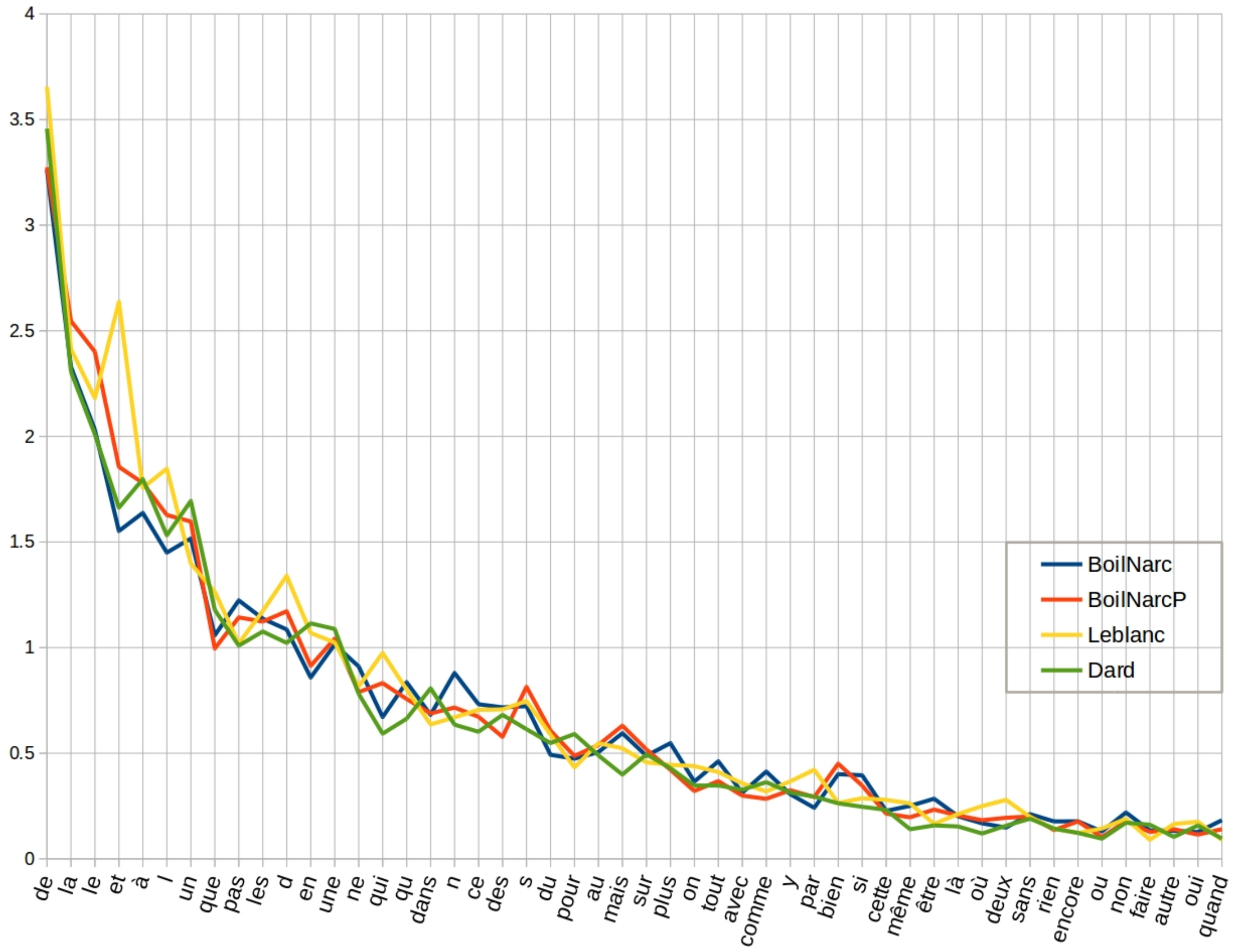
	A	B	C	D	E	F	G
1		**Leblanc_1934=rp046.txt**	**Leblanc_1935=rp039.txt**	**BoilNarcP_1974=rp260.txt**	**BoilNarcP_1975=rp261.txt**	**BoilNarc_1955=rp253.txt**	**BoilNarcP_1979=rp263.txt**
2	**de**	0.0364230093	0.0360980047	0.0324750152	0.0342983665	0.0345434543	0.0319564549
3	**la**	0.0262979924	0.0222841226	0.0270066417	0.0247315871	0.0215821582	0.0231113647
4	**il**	0.0224141596	0.023743202	0.0273628193	0.0256715545	0.0080008001	0.026798648
5	**le**	0.0210809035	0.0223030716	0.0204487838	0.025232903	0.0174817482	0.0233747421
6	**et**	0.0250420265	0.0272677316	0.0184583796	0.0173998412	0.0159615962	0.0181291427
7	**à**	0.0182791336	0.0168646846	0.0173479436	0.0184024732	0.0173617362	0.0173390106
8	**l**	0.0213707418	0.018986982	0.0173688952	0.0155825709	0.0123012301	0.016658619
9	**je**	0.0082893745	0.0103461998	0.0122357477	0.0144337219	0.0333433343	0.0113471753
10	**un**	0.0146464939	0.0137570348	0.0151899265	0.0168776371	0.0147414741	0.0146613406
11	**d**	0.0141634301	0.0144581509	0.0096586981	0.0118227013	0.0117211721	0.0125982178
12	**pas**	0.0111297896	0.0115020939	0.0115443441	0.0118644776	0.0114611461	0.0115886045
13	**que**	0.0118640465	0.0134349004	0.0103920049	0.0103396416	0.0118211821	0.0100302884
14	**elle**	0.0095260178	0.0068974665	0.0086530202	0.0044074028	0.0195619562	0.0104473026
15	**les**	0.0097192433	0.0087544767	0.0128223931	0.0109662865	0.0115811581	0.0100302884
16	**une**	0.0102216297	0.0092282038	0.0096796497	0.0109453983	0.0101210121	0.0101839252
17	**est**	0.0121538848	0.0108388759	0.0110624568	0.0099845428	0.0086408641	0.0122909442
18	**vous**	0.0107433386	0.0101946071	0.0126547801	0.0154990183	0.0083208321	0.0119178263
19	**en**	0.0103568875	0.0107441305	0.008883488	0.0090445753	0.0080208021	0.0097010667
20	**ne**	0.0083666647	0.0092471529	0.0079616166	0.0081046079	0.0097209721	0.0079671656
21	**qui**	0.010337565	0.0099293199	0.0074587777	0.00887747	0.0072007201	0.0086695053
22	**n**	0.0072846019	0.0079396661	0.0073121163	0.0068722062	0.0079807981	0.0075062552
23	**qu**	0.0080381814	0.0077880734	0.0079825682	0.0065171074	0.0064006401	0.00803301
24	**s**	0.0074198597	0.0069543138	0.008275891	0.0086477002	0.004760476	0.0075282033
25	**était**	0.0042896064	0.0053057435	0.0078568585	0.0064335547	0.0056605661	0.0081427505
26	**se**	0.0069367959	0.0068406193	0.0084016007	0.0080210553	0.00410041	0.0077696326
27	**ce**	0.0073618921	0.0075038372	0.006055019	0.0067886535	0.0065606561	0.0069795005
28	**dans**	0.0069174734	0.0075796335	0.0080873264	0.0060784559	0.0073807381	0.0061674202
29	**des**	0.0062798292	0.0058363178	0.0063902449	0.006976647	0.0064606461	0.0047188447
30	**avait**	0.0046760574	0.0047751691	0.0061178738	0.0048251661	0.004560456	0.0067380712
31	**mais**	0.004907928	0.0059689614	0.0064530998	0.0050967122	0.0056005601	0.0071989816

Vector similarity / distance



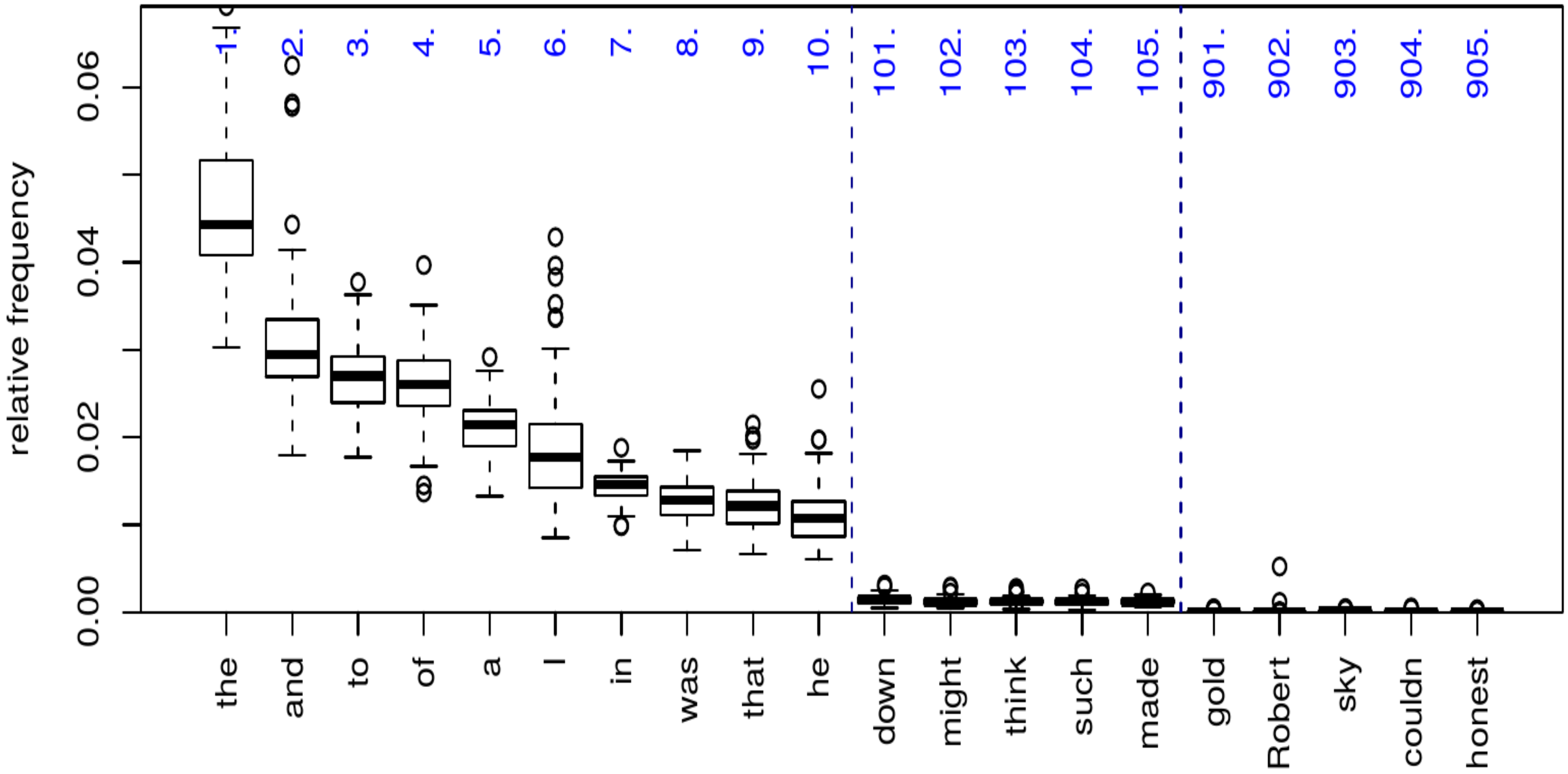
(Two texts, A and B; two words on x and y axes)

Frequency vectors plotted



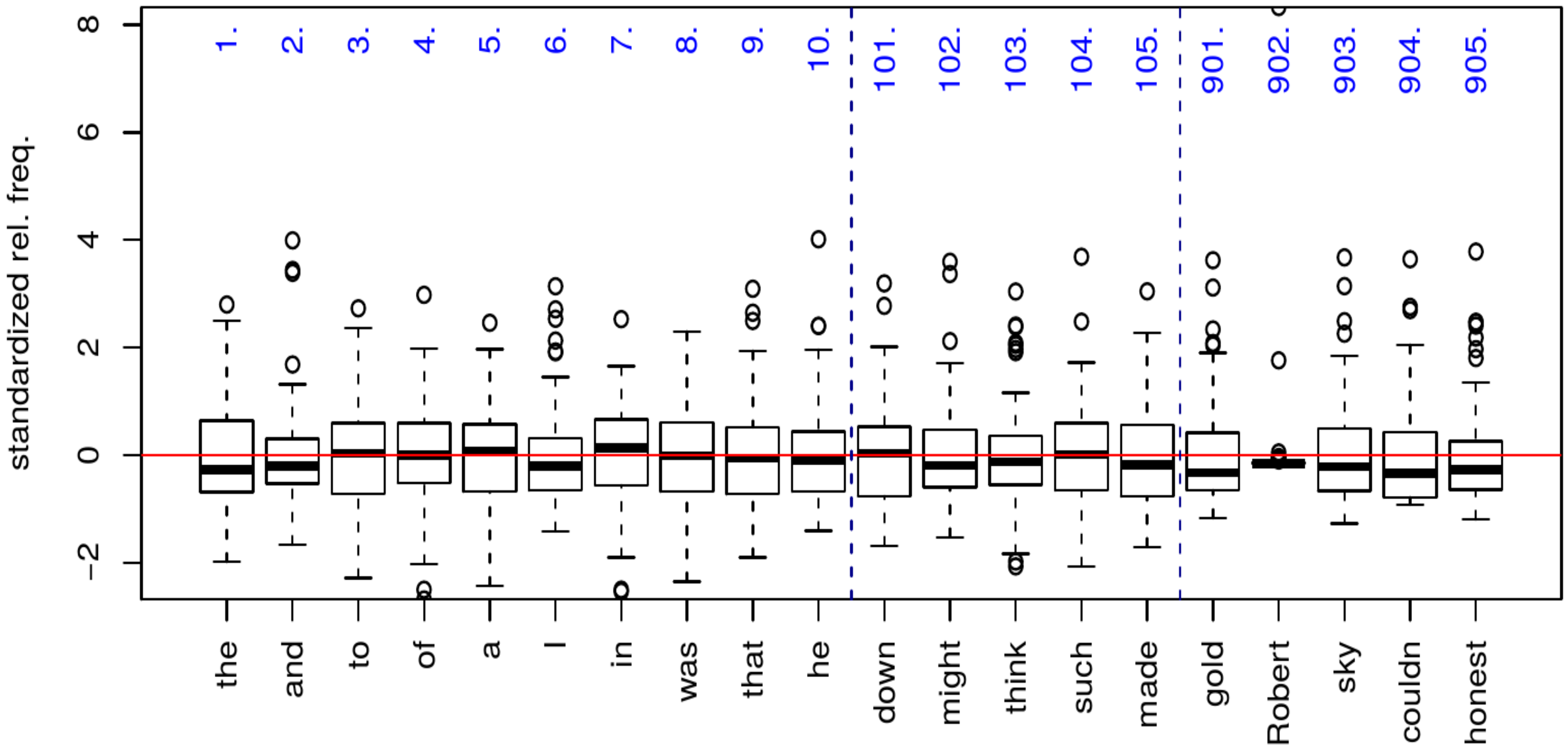
(One text for three/four different authors)

Distribution of relative frequencies



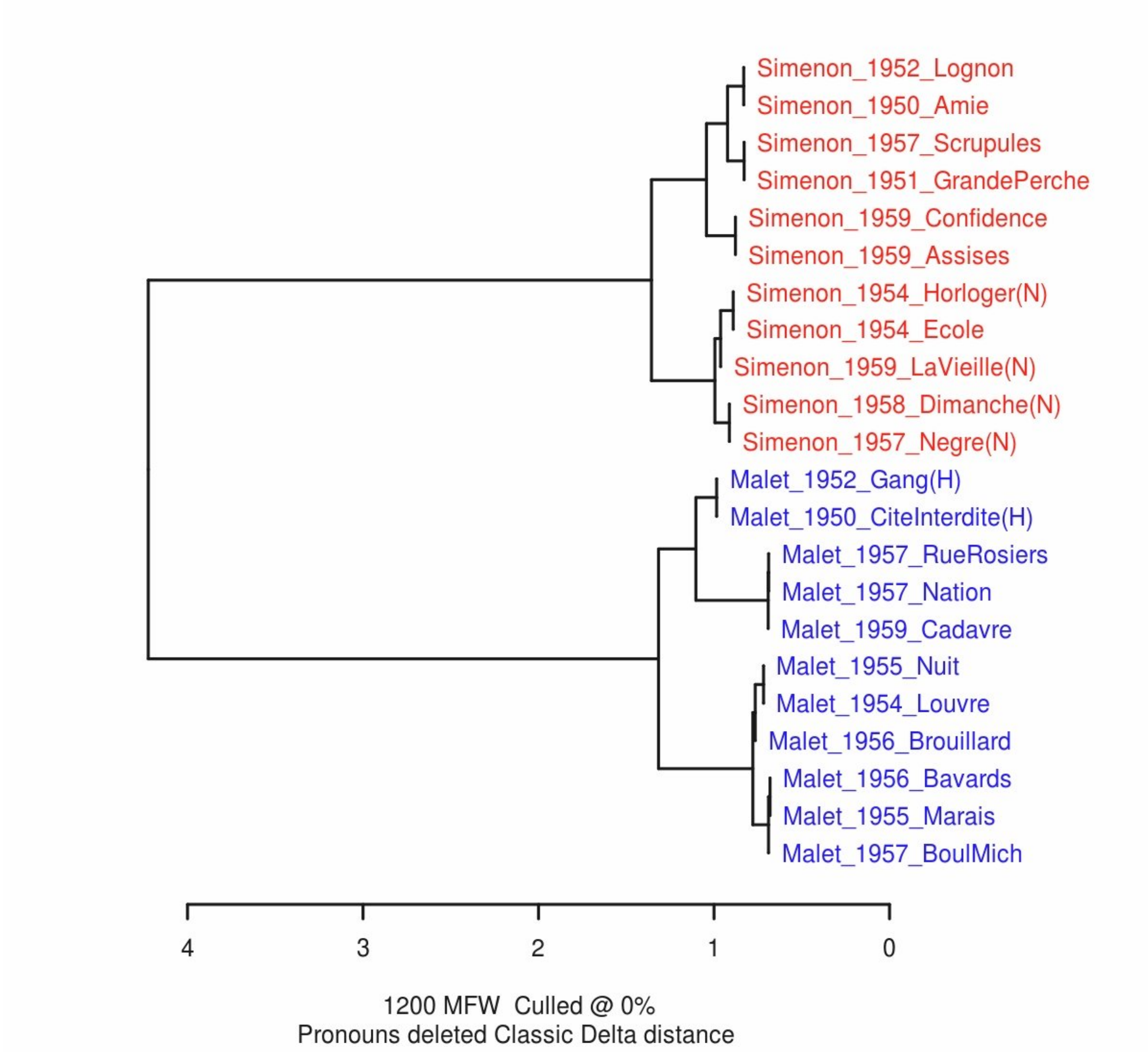
(High-frequency words dominate; graph by Stefan Evert)

Distribution of standardized frequencies



(This is what Burrows' Delta does; graph by Stefan Evert)

Dendrogram


Simenon_1952_Lognon
Simenon_1950_Amie
Simenon_1957_Scrupules
Simenon_1951_GrandePerche
Simenon_1959_Confidence
Simenon_1959_Assises
Simenon_1954_Horloger(N)
Simenon_1954_Ecole
Simenon_1959_LaVieille(N)
Simenon_1958_Dimanche(N)
Simenon_1957_Negre(N)
Malet_1952_Gang(H)
Malet_1950_CiteInterdite(H)
Malet_1957_RueRosiers
Malet_1957_Nation
Malet_1959_Cadavre
Malet_1955_Nuit
Malet_1954_Louvre
Malet_1956_Brouillard
Malet_1956_Bavards
Malet_1955_Marais
Malet_1957_BoulMich
4
3
2
1
0
1200 MFW Culled @ 0%
Pronouns deleted Classic Delta distance

(Based on the similarity matrix)

What does such similarity tell us?

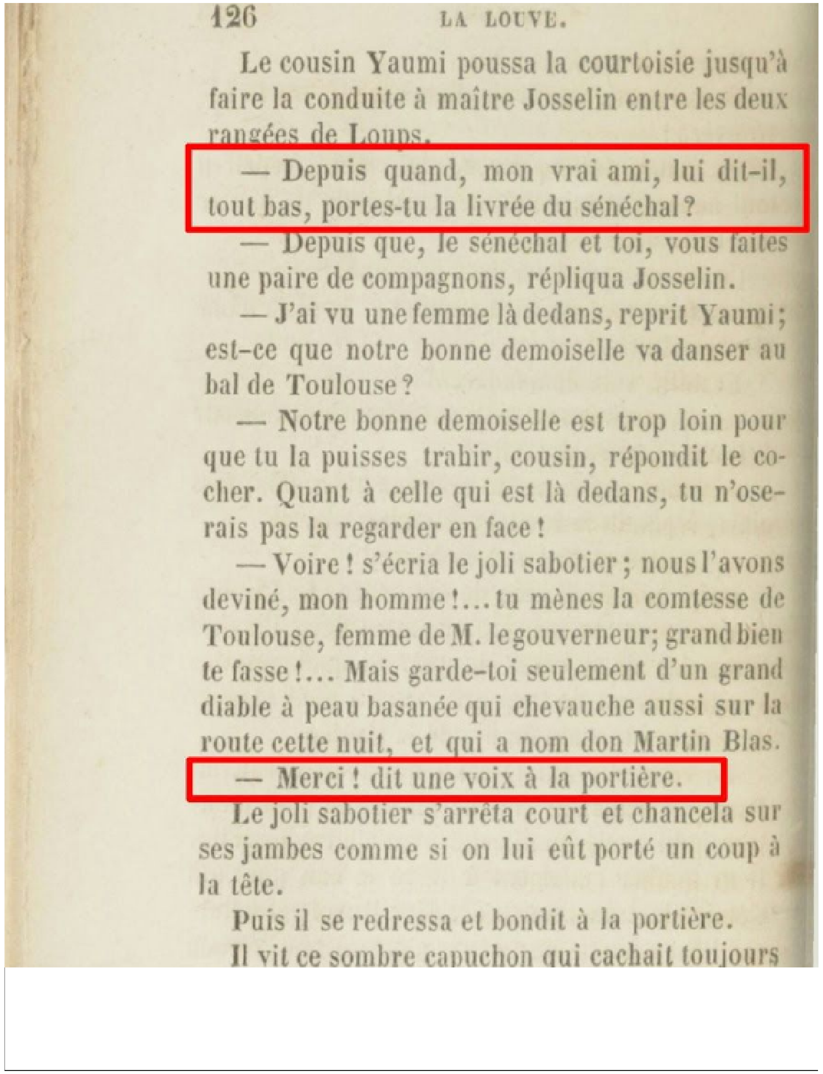
3. Machine Learning in Computational Narratology

(Example: Automatic Recognition of Direct Speech, with Daniel Schlör and Stefanie Popp)

Automatic Recognition of Direct Speech

- Computational Narratology
- Relies on annotation and Machine Learning
- Necessitates feature generation
- Derives a higher-order feature from surface / lower-order features

What is the problem?



126 LA LOUVE.

Le cousin Yaumi poussa la courtoisie jusqu'à faire la conduite à maître Josselin entre les deux rangées de Loups.

— Depuis quand, mon vrai ami, lui dit-il, tout bas, portes-tu la livrée du sénéchal ?

— Depuis que, le sénéchal et toi, vous faites une paire de compagnons, répliqua Josselin.

— J'ai vu une femme là dedans, reprit Yaumi ; est-ce que notre bonne demoiselle va danser au bal de Toulouse ?

— Notre bonne demoiselle est trop loin pour que tu la puisses trahir, cousin, répondit le cocher. Quant à celle qui est là dedans, tu n'oserais pas la regarder en face !

— Voire ! s'écria le joli sabotier ; nous l'avons deviné, mon homme !... tu mènes la comtesse de Toulouse, femme de M. le gouverneur ; grand bien te fasse !... Mais garde-toi seulement d'un grand diable à peau basanée qui chevauche aussi sur la route cette nuit, et qui a nom don Martin Blas.

— Merci ! dit une voix à la portière.

Le joli sabotier s'arrêta court et chancela sur ses jambes comme si on lui eût porté un coup à la tête.

Puis il se redressa et bondit à la portière.

Il vit ce sombre capuchon qui cachait toujours

The corpus used

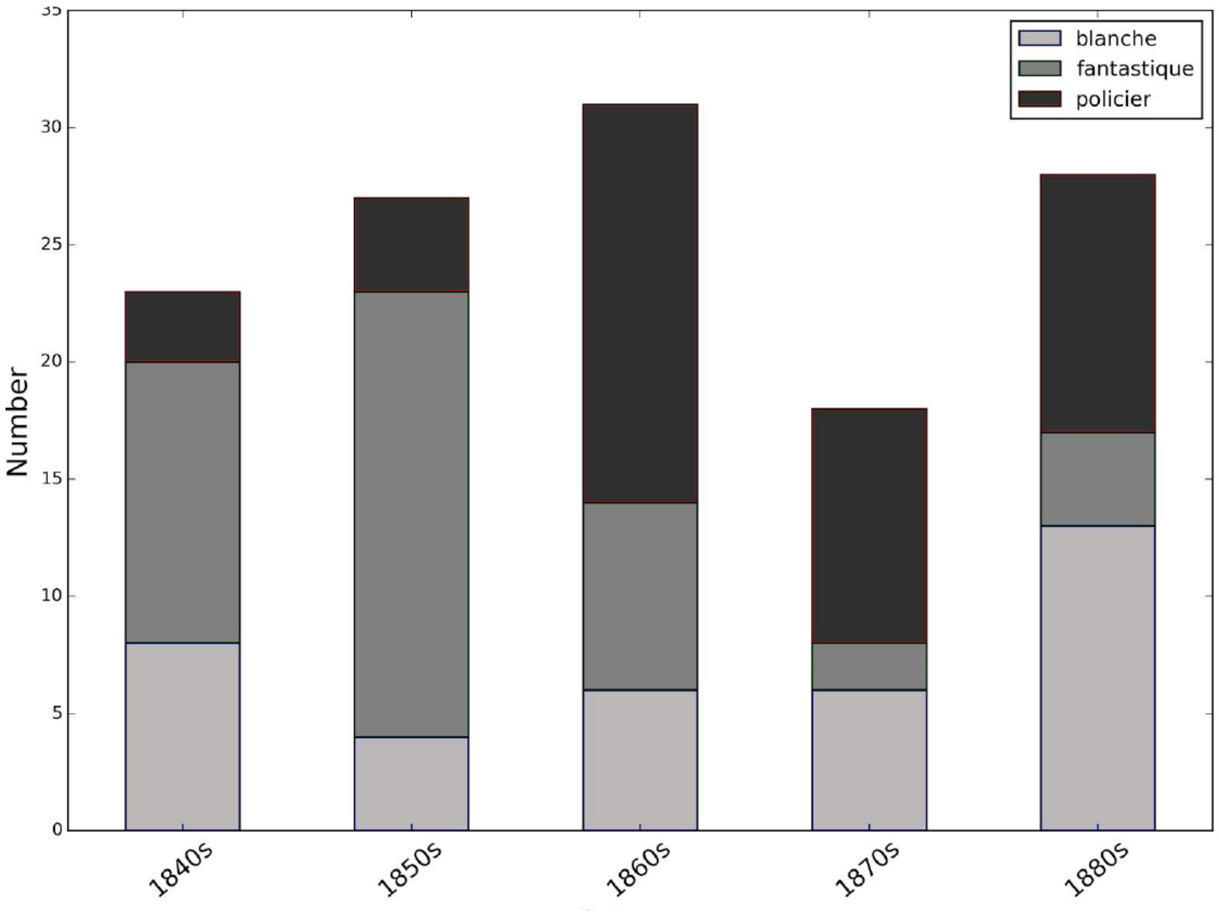


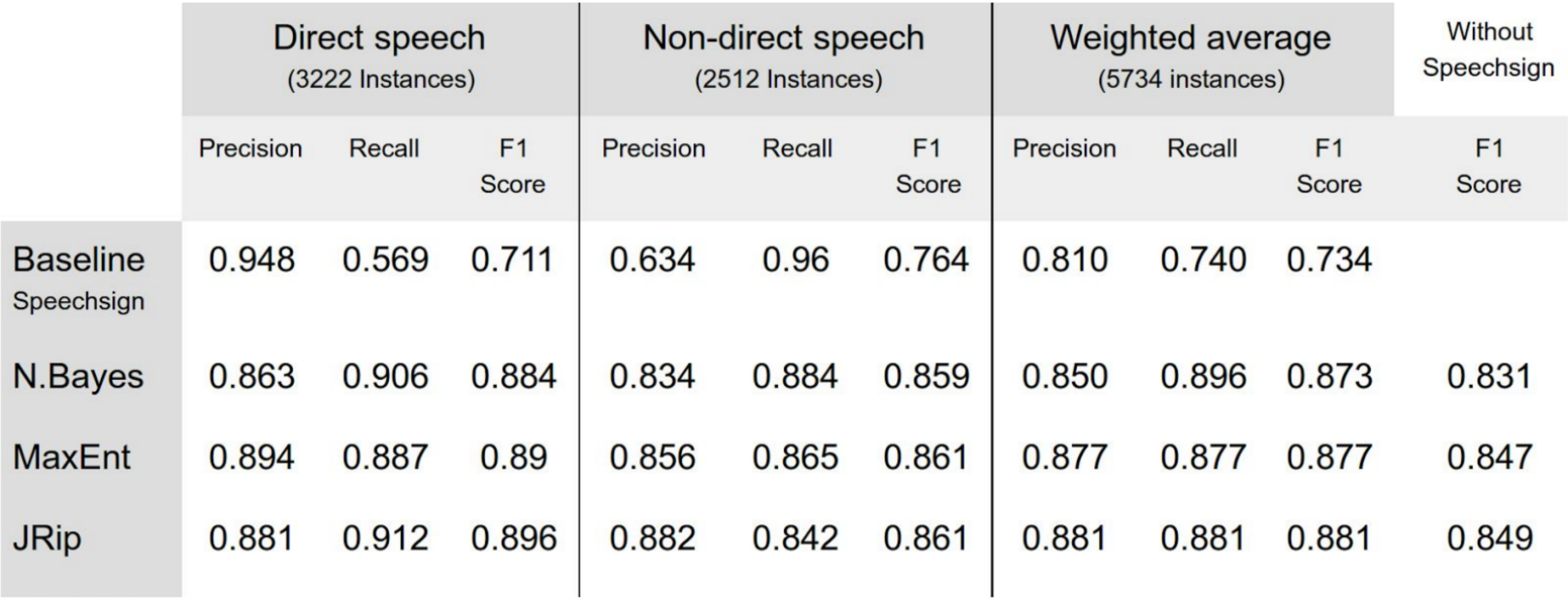
Figure 2: Distribution of novels per subgenre and decade.

127 French Novels, 1840-1890

The method

Feature generation (81 features)

Performance



	Direct speech (3222 Instances)			Non-direct speech (2512 Instances)			Weighted average (5734 instances)			Without Speechsign
	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score	F1 Score
Baseline Speechsign	0.948	0.569	0.711	0.634	0.96	0.764	0.810	0.740	0.734	
N.Bayes	0.863	0.906	0.884	0.834	0.884	0.859	0.850	0.896	0.873	0.831
MaxEnt	0.894	0.887	0.89	0.856	0.865	0.861	0.877	0.877	0.877	0.847
JRip	0.881	0.912	0.896	0.882	0.842	0.861	0.881	0.881	0.881	0.849

Random Forests: F1-score 0.938

Results: Subgenres

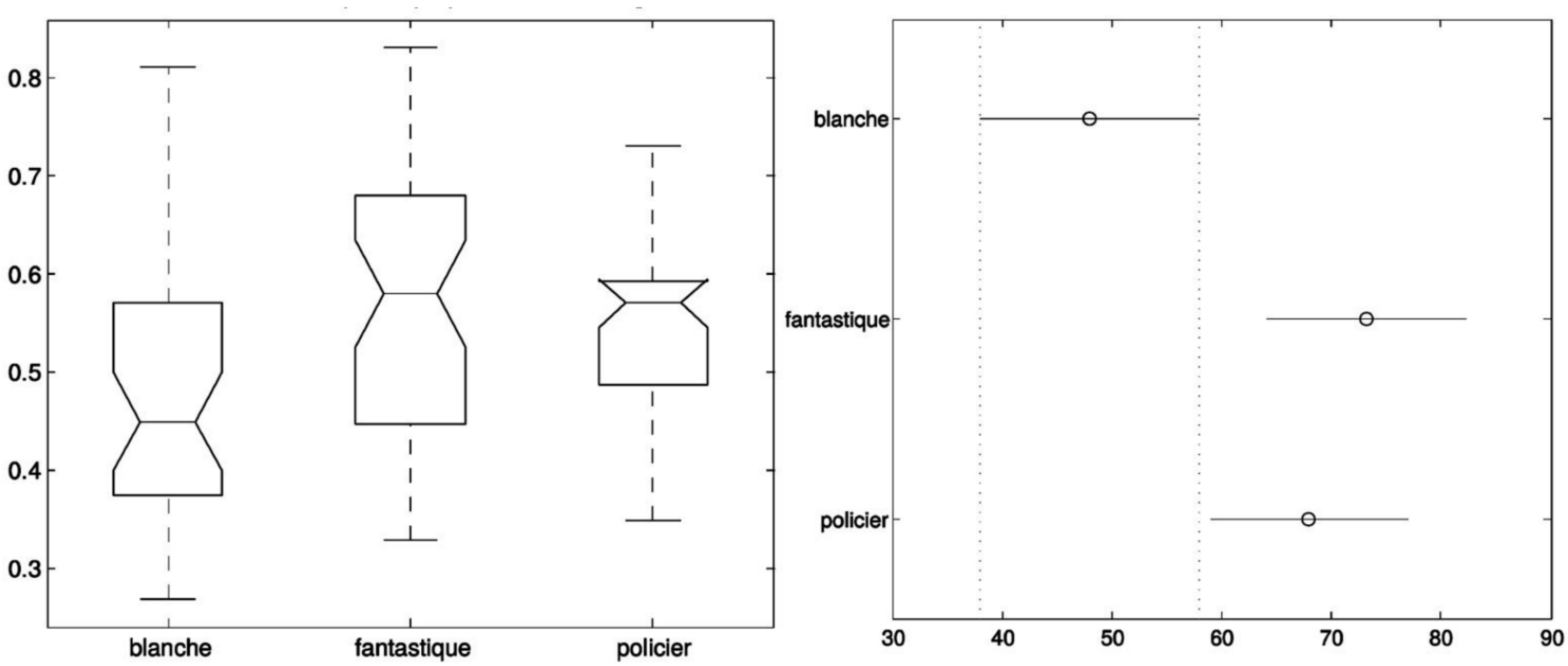


Figure 4: Distribution (left) and significance (right)

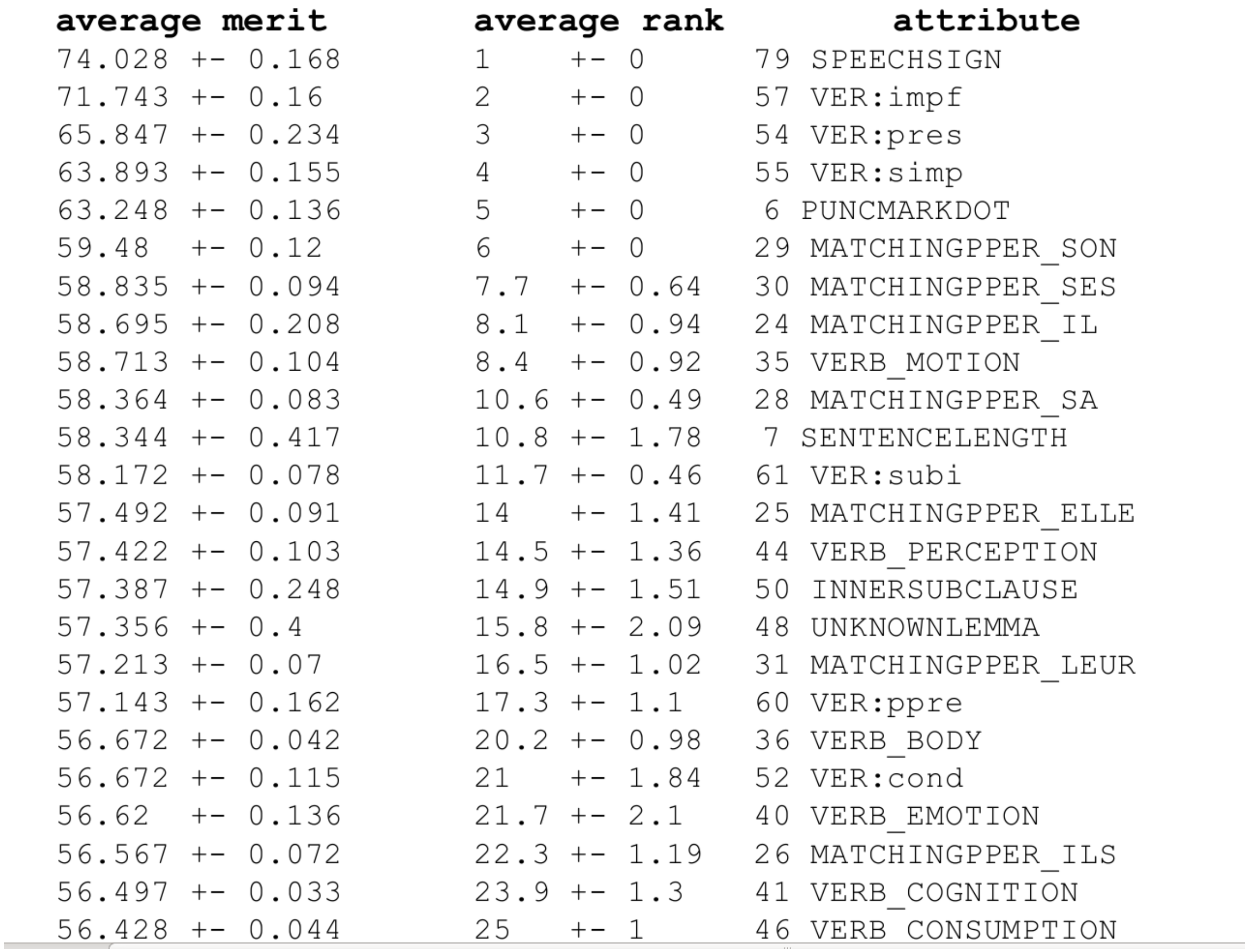
of direct to non-direct speech ratios across three subgenres

Popular vs. literary subgenres

What is it useful for?

- Investigate subgenre preferences and long-term evolution
- Investigate differences between different charachter's styles
- Separate narrator and character speech for stylometry
- Identify narrator speech and further classify it (narrative, descriptive)

Top Features



```
average merit         average rank        attribute
74.028 +- 0.168       1    +- 0       79 SPEECHSIGN
71.743 +- 0.16        2    +- 0       57 VER:impf
65.847 +- 0.234       3    +- 0       54 VER:pres
63.893 +- 0.155       4    +- 0       55 VER:simp
63.248 +- 0.136       5    +- 0        6 PUNCMARKDOT
59.48  +- 0.12        6    +- 0       29 MATCHINGPPER_SON
58.835 +- 0.094       7.7  +- 0.64    30 MATCHINGPPER_SES
58.695 +- 0.208       8.1  +- 0.94    24 MATCHINGPPER_IL
58.713 +- 0.104       8.4  +- 0.92    35 VERB_MOTION
58.364 +- 0.083       10.6 +- 0.49    28 MATCHINGPPER_SA
58.344 +- 0.417       10.8 +- 1.78     7 SENTENCELENGTH
58.172 +- 0.078       11.7 +- 0.46    61 VER:subi
57.492 +- 0.091       14   +- 1.41    25 MATCHINGPPER_ELLE
57.422 +- 0.103       14.5 +- 1.36    44 VERB_PERCEPTION
57.387 +- 0.248       14.9 +- 1.51    50 INNERSUBCLAUSE
57.356 +- 0.4         15.8 +- 2.09    48 UNKNOWNLEMMA
57.213 +- 0.07        16.5 +- 1.02    31 MATCHINGPPER_LEUR
57.143 +- 0.162       17.3 +- 1.1     60 VER:ppre
56.672 +- 0.042       20.2 +- 0.98    36 VERB_BODY
56.672 +- 0.115       21   +- 1.84    52 VER:cond
56.62  +- 0.136       21.7 +- 2.1     40 VERB_EMOTION
56.567 +- 0.072       22.3 +- 1.19    26 MATCHINGPPER_ILS
56.497 +- 0.033       23.9 +- 1.3     41 VERB_COGNITION
56.428 +- 0.044       25   +- 1       46 VERB_CONSUMPTION
```

long hyphen, verb tense, personal pronouns

4. Thematic Analysis with Topic Modeling

Basic Idea

Words, Topics, Documents

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

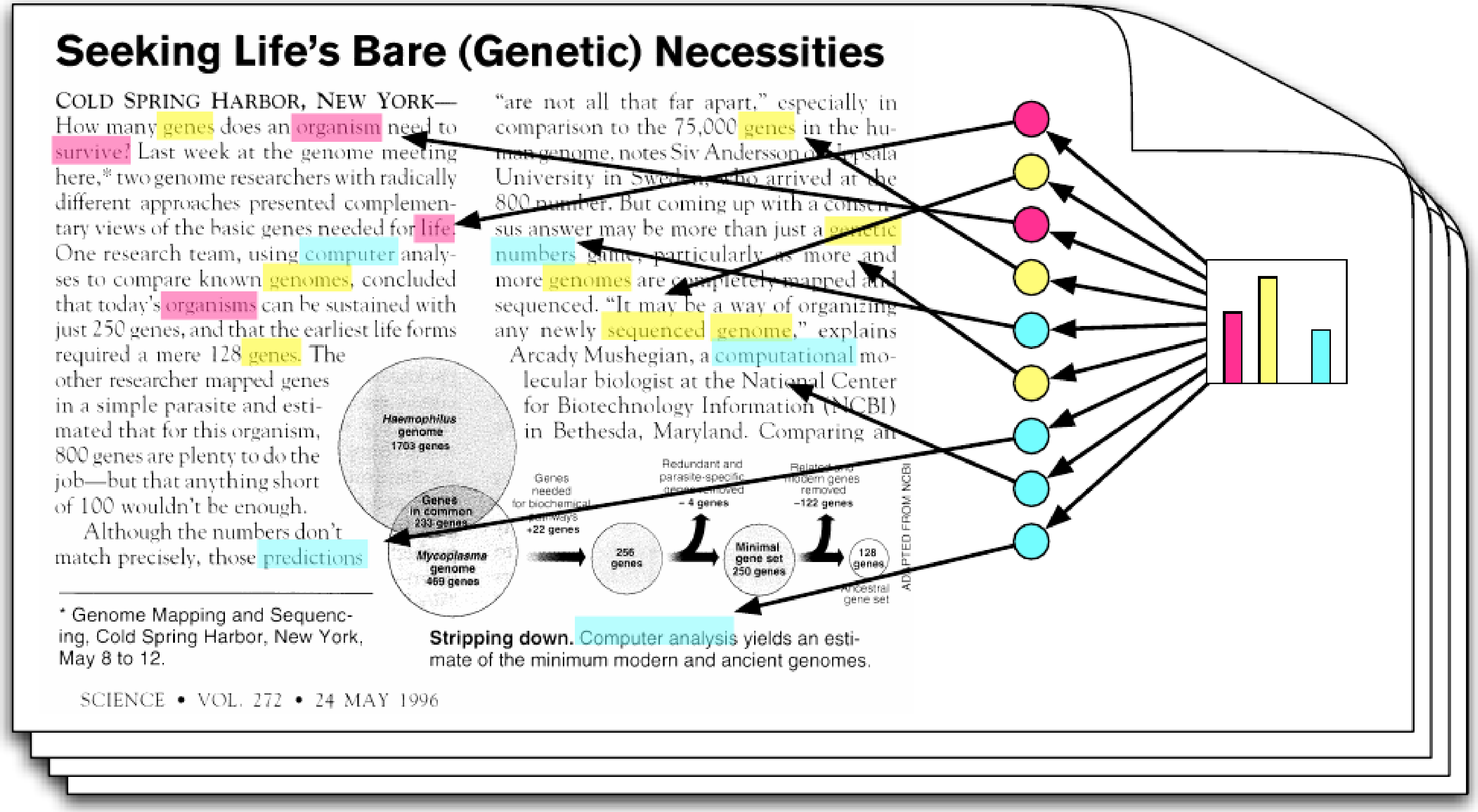
life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Topic proportions and assignments



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

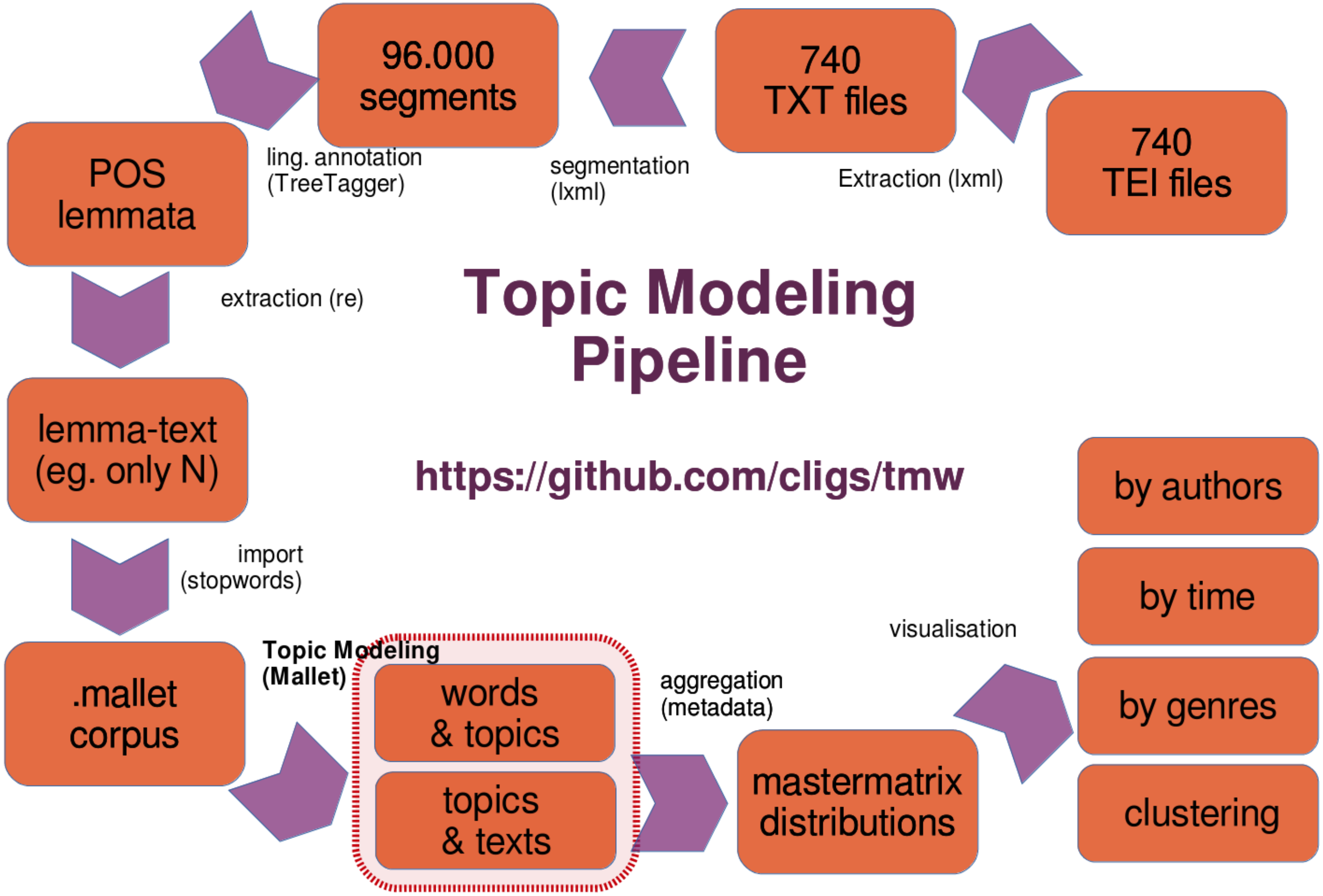
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

(David Blei, "Probabilistic Topic Models", 2012)

Preprocessing, Modeling, Postprocessing



Topic Modeling Pipeline

https://github.com/cligs/tmw

740 TEI files

Extraction (lxml)

740 TXT files

segmentation (lxml)

96.000 segments

ling. annotation (TreeTagger)

POS lemmata

extraction (re)

lemma-text (eg. only N)

import (stopwords)

.mallet corpus

Topic Modeling (Mallet)

words & topics

topics & texts

aggregation (metadata)

mastermatrix distributions

visualisation

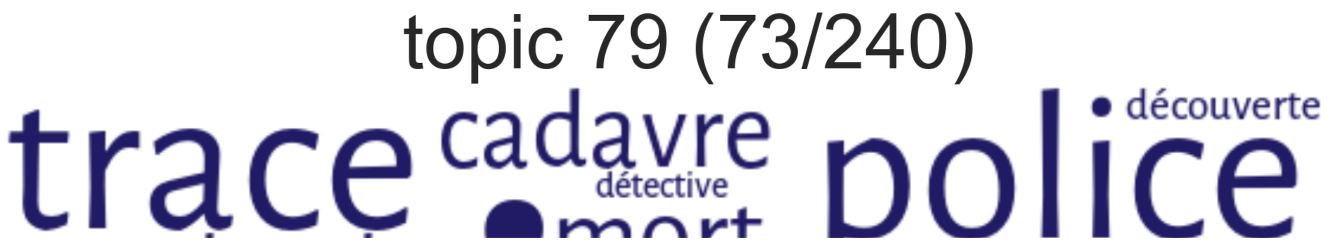
by authors

by time

by genres

clustering

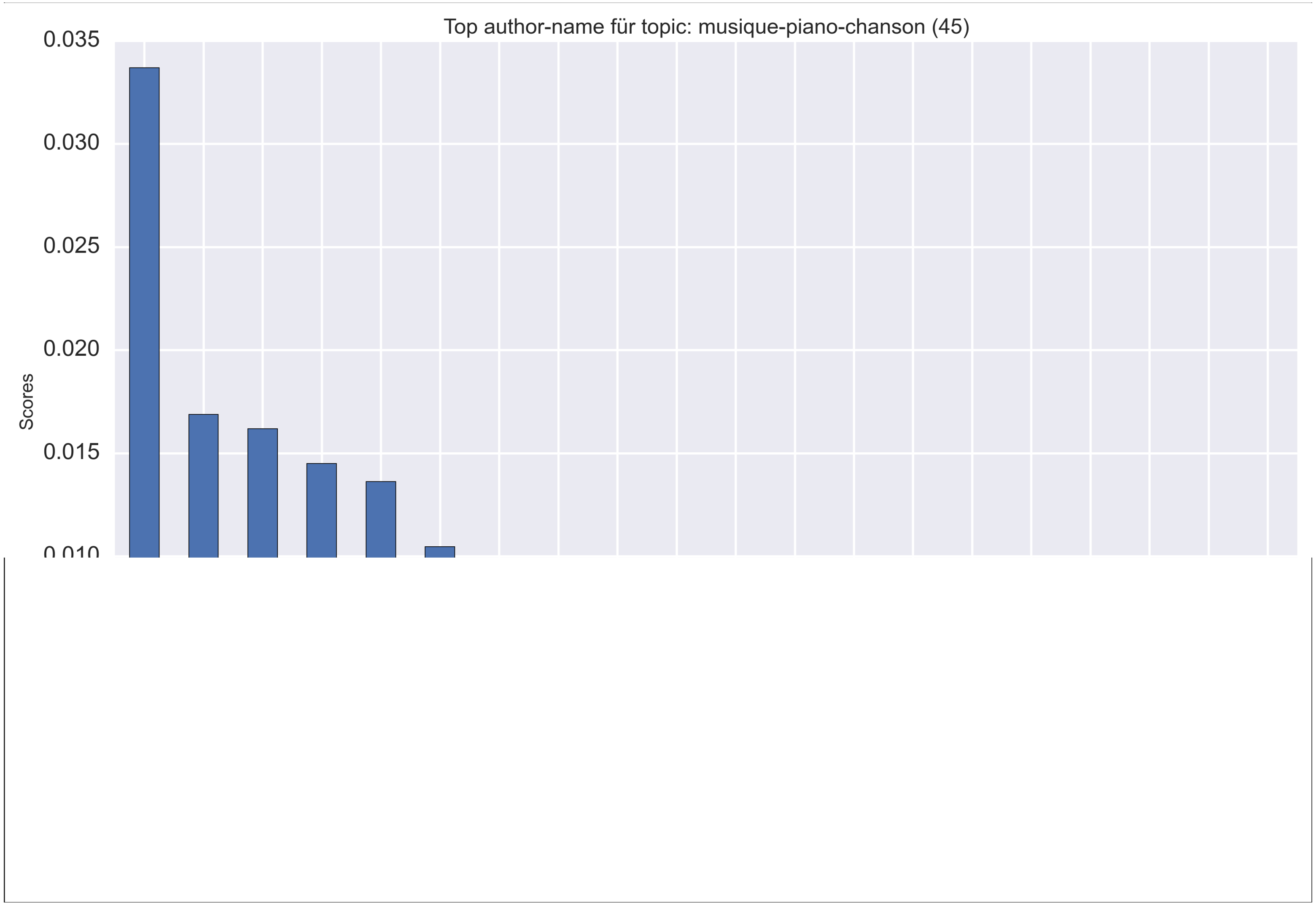
Topics: Crime



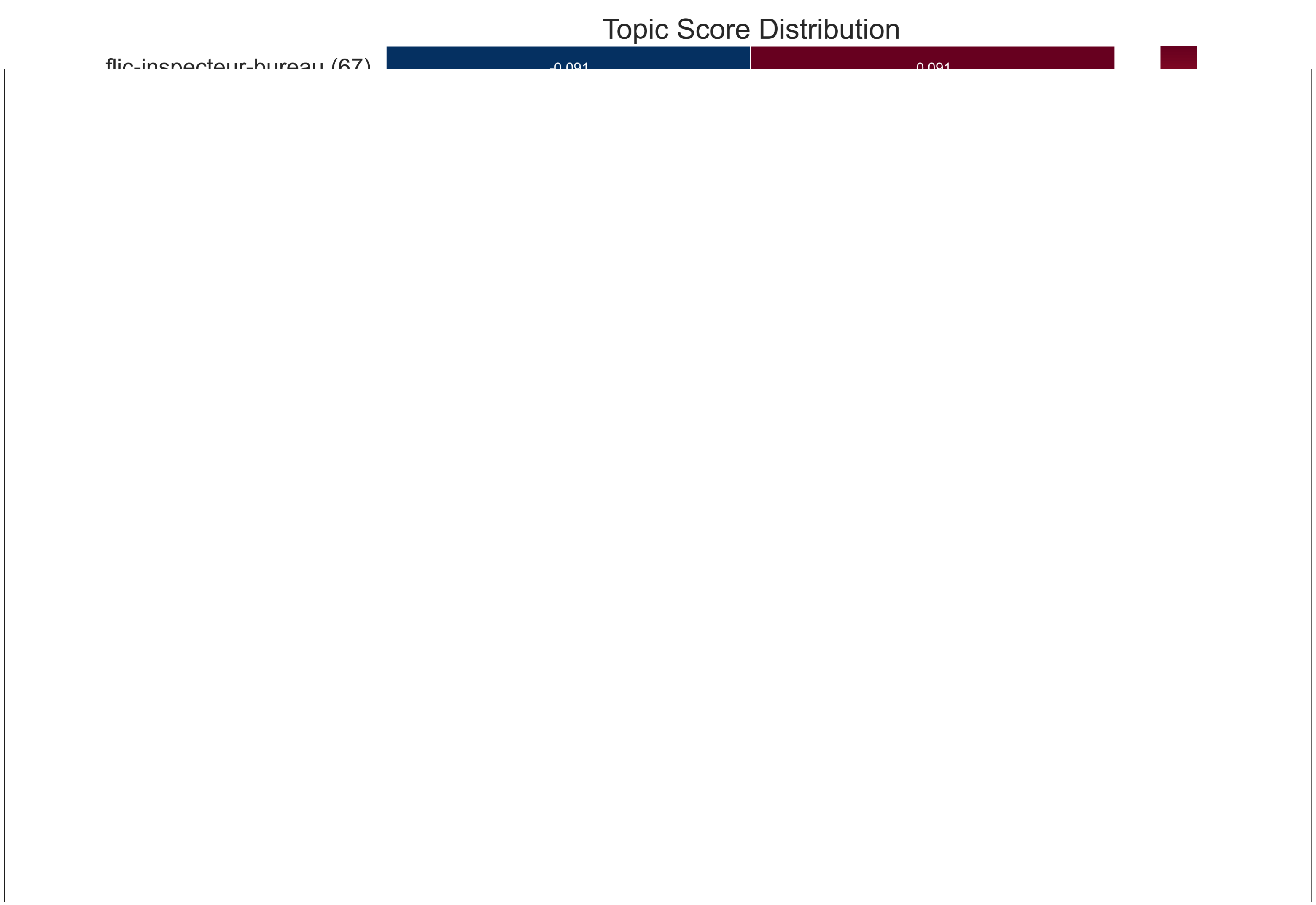
Topics: Music

topic 45 (102/240)
œuvre
clavier
artiste
mesure
mélodie

Top-Authors for Music



Crime fiction vs. non-crime fiction



Topics over time

What is it good for?

Conclusion

Conclusion

Further Reading

- Shillingsburg, P. (2006). *From Gutenberg to Google. Electronic Representations of Literary Texts*. Cambridge: Cambridge Univ. Press.
- Alpaydin, E. (2010). *Introduction to Machine Learning*. 2nd ed. Cambridge, Mass: MIT Press.
- Ramsay, S. (2011). *Reading Machines : Toward an Algorithmic Criticism*. Urbana Ill.: University of Illinois Press.
- *Doing Digital Humanities Bibliography*: https://www.zotero.org/groups/doing_digital_humanities_-_a_dariah_bibliography

Thank you!

Christof Schöch, 2016

http://www.christof-schoech.de
