Comparison of Text Groups using Measures of Distinctiveness

Christof Schöch (Univ. of Trier)
with Daniel Schlör, Albin Zehe, Henning Gebhard, Andreas Hotho, Cora Rok, Keli Du, Julia Dudar, Julian Schröter.

Neo-Latin Studies and Digital Humanities | Bonn, April 14-16, 2021

Universität Trier

Overview

1. Introduction: What is distinctiveness?
2. Understanding Zeta
3. Variants and Evaluation
4. Application: French Novel
5. Current Work
6. Conclusion

1. Introduction: What is Distinctiveness?

Foundations

- Contrastive / comparative analysis is widespread
- Numerous measures of distinctiveness ("keyness")
- Many tools have implementations
  - Antconc
  - WordCruncher
  - TXM
  - stylo
  - Intelligent Archive
  - etc.

An intuition for distinctiveness

An intuition for distinctiveness

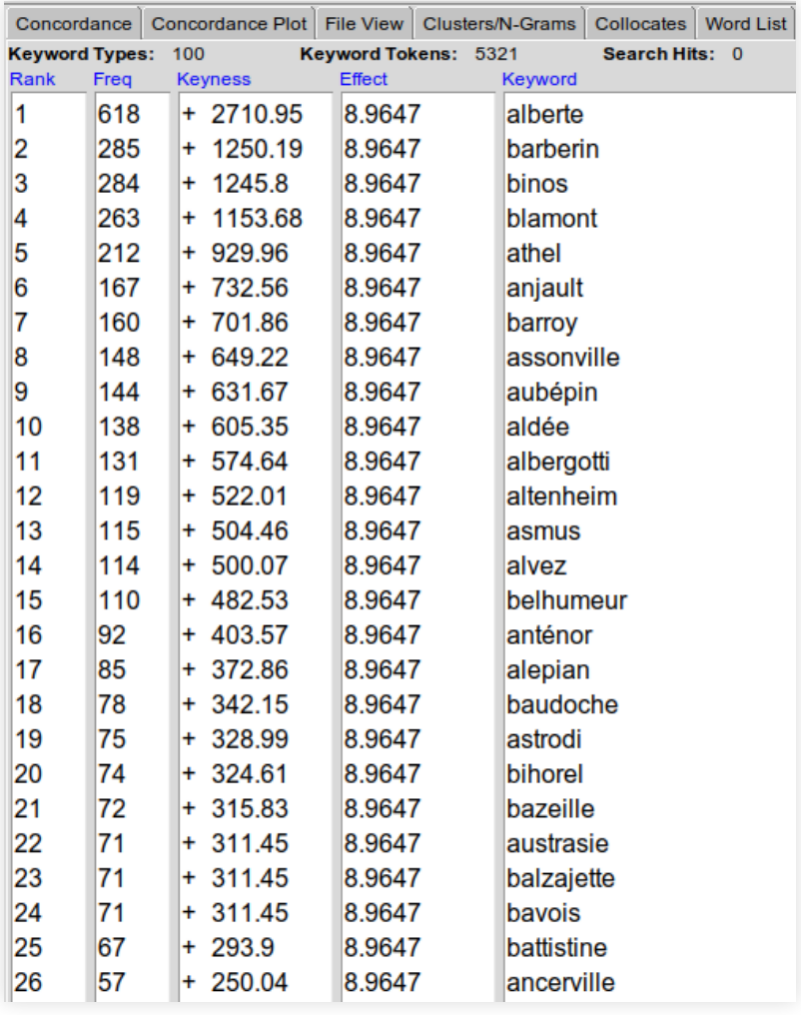
German "Apfelschorle"

What is distinctiveness (conceptually)?

Some properties of distinctive features

- they are typical for a group
- characteristic of a group
- express aboutness
- have discriminatory power
- would be salient or surprising in the opposite group

Log-likelihood in Antconc



Concordance | Concordance Plot | File View | Clusters/N-Grams | Collocates | Word List

Keyword Types: 100 **Keyword Tokens:** 5321 **Search Hits:** 0

Rank	Freq	Keyness	Effect	Keyword
1	618	+ 2710.95	8.9647	alberte
2	285	+ 1250.19	8.9647	barberin
3	284	+ 1245.8	8.9647	binos
4	263	+ 1153.68	8.9647	blamont
5	212	+ 929.96	8.9647	athel
6	167	+ 732.56	8.9647	anjault
7	160	+ 701.86	8.9647	barroy
8	148	+ 649.22	8.9647	assonville
9	144	+ 631.67	8.9647	aubépin
10	138	+ 605.35	8.9647	aldée
11	131	+ 574.64	8.9647	albergotti
12	119	+ 522.01	8.9647	altenheim
13	115	+ 504.46	8.9647	asmus
14	114	+ 500.07	8.9647	alvez
15	110	+ 482.53	8.9647	belhumeur
16	92	+ 403.57	8.9647	anténor
17	85	+ 372.86	8.9647	alepian
18	78	+ 342.15	8.9647	baudoche
19	75	+ 328.99	8.9647	astrodi
20	74	+ 324.61	8.9647	bihorel
21	72	+ 315.83	8.9647	bazeille
22	71	+ 311.45	8.9647	austrasie
23	71	+ 311.45	8.9647	balzajette
24	71	+ 311.45	8.9647	bavois
25	67	+ 293.9	8.9647	battistine
26	57	+ 250.04	8.9647	ancerville

What is distinctiveness (statistically)?

- Relies on the comparison of two groups
- Calculates a score for each feature
- Simple frequency is not enough (= typical)
- Rather: comparatively high/low frequency (= distinctive)
- Not: entirely mutually exclusive (= purely discriminant)
- And: not just frequency, but also distribution / dispersion

Four types of measures

- Based on frequency (e.g. log-likelihood ratio)
- Based on distribution (e.g. t-Test)
- Based on dispersion (e.g. Zeta)
- Based on Machine Learning (e.g. weights from linear SVM)

- Recommended readings
  - Lijffijt et al., "Significance testing of word frequencies in corpora", 2014
  - Gries, "A new approach to (key) keywords analysis", 2021

2. Understanding Zeta

Zeta

- A measure of distinctiveness developed in Computational Literary Studies
- Based on the dispersion of features (rather than pure frequency)
- Bias towards medium-frequency content words: high interpretability of results
- Ignored by virtually all relevant work in CL ;-)

Previous work

- Proposed by John Burrows (2007) in the context of authorship attribution
- Studies by Hugh Craig (2009), David Hoover (2010), Rizvi (2018)
- Our work so far:
  - a Python implementation (pyzeta)
  - an application to literary subgenres (Schöch 2018)
  - an evaluation study (Schöch et al. 2018)
- Currently: "Zeta and Company"

Zeta and Company

- Zeta and Company: Measures of Distinctiveness for Computational Literary Studies (2020-2023); see: https://zeta-project.eu
- Project team: Julia Dudar (CL), Cora Rok (LS), Keli Du (CLS).
- Part of the DFG Priority Programme Computational Literary Studies (SPP 2207); see: https://dfg-spp-cls.github.io/
- Key objective: Model, implement, evaluate, and use of various measures of ‘keyness’ or ‘distinctiveness’
- Further characteristics
  - Focus on comparison of text corpora on the lexical level
  - Building bridges between IR, CL and CLS communities

Zeta: Basics

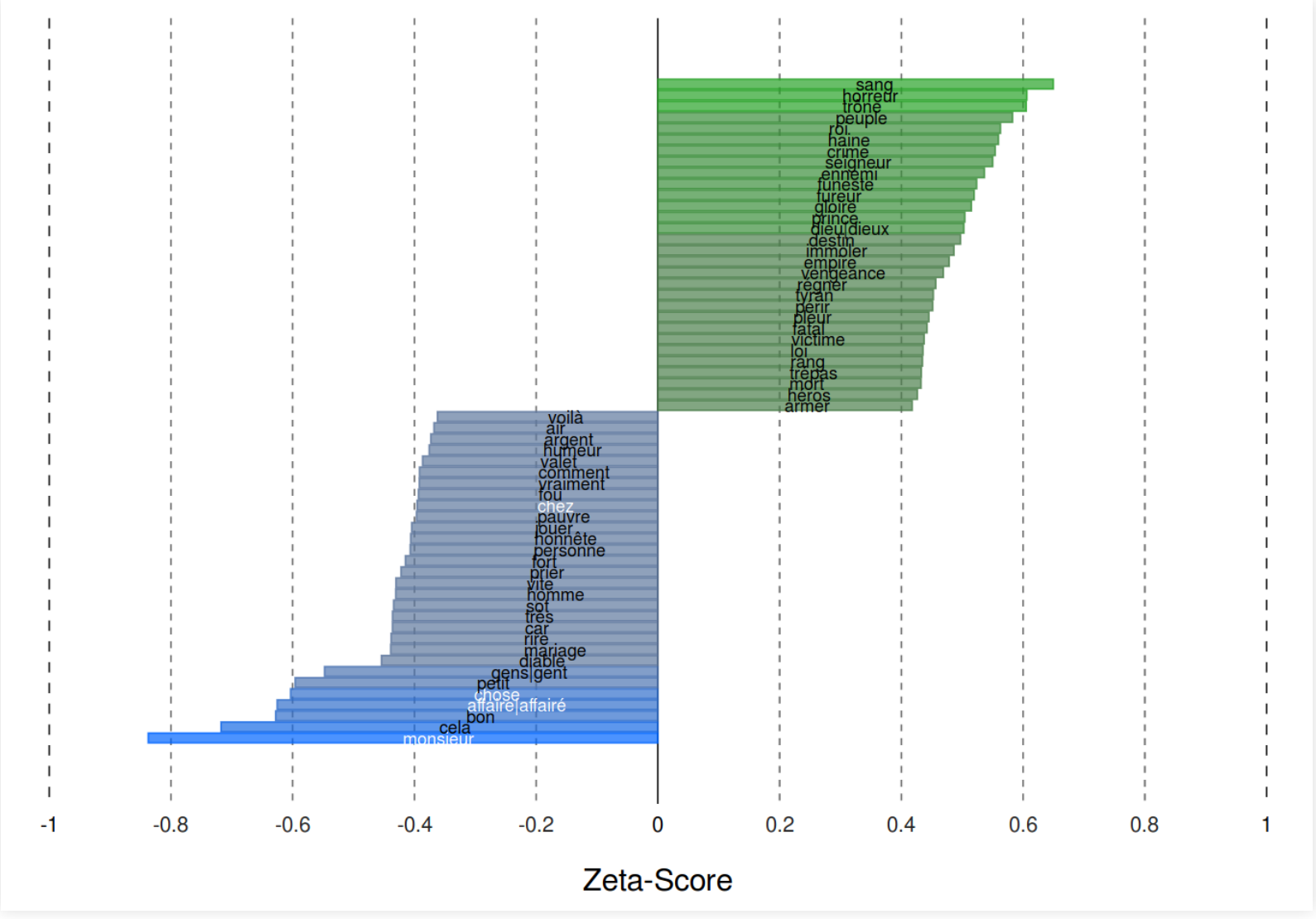
- two groups of documents G1 and G2
- each document is split into m segments of n words
- sp_i = segment proportion of word type i
- segment proportion: the proportion of segments in one group that contain at least one instance of the word type
- calculated for G1 and G2 separatedly

Zeta: Calculation

$$\mathbf{Zeta_j = sp_j(G1) - sp_j(G2)}$$

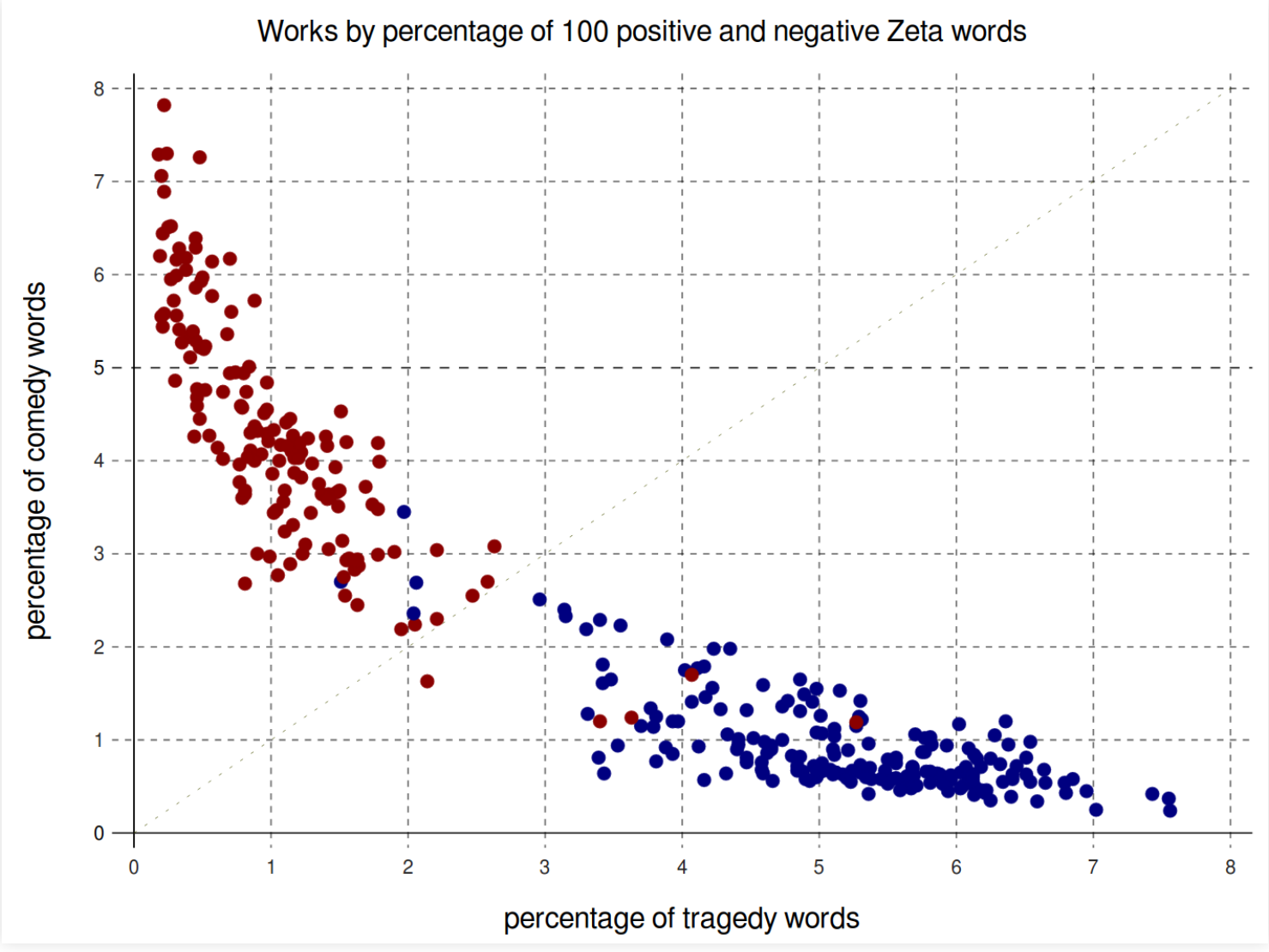
- A simple subtraction of the segment proportions
- Calculated for each word type, sorted by Zeta
- Result: list of distinctive words

Illustration: Tragedies vs comedies



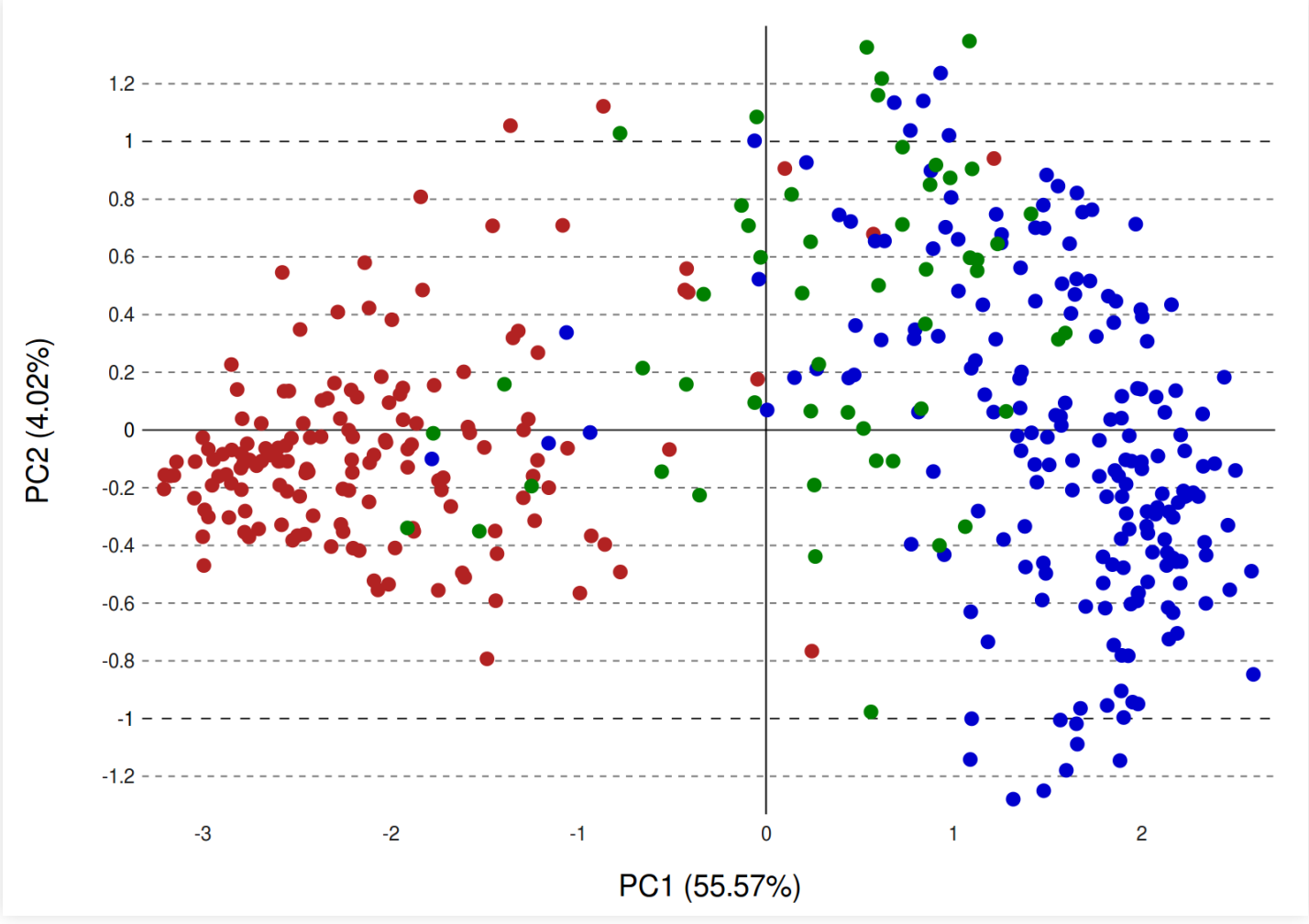
Comedy and tragedy words

Proportions of Zeta words



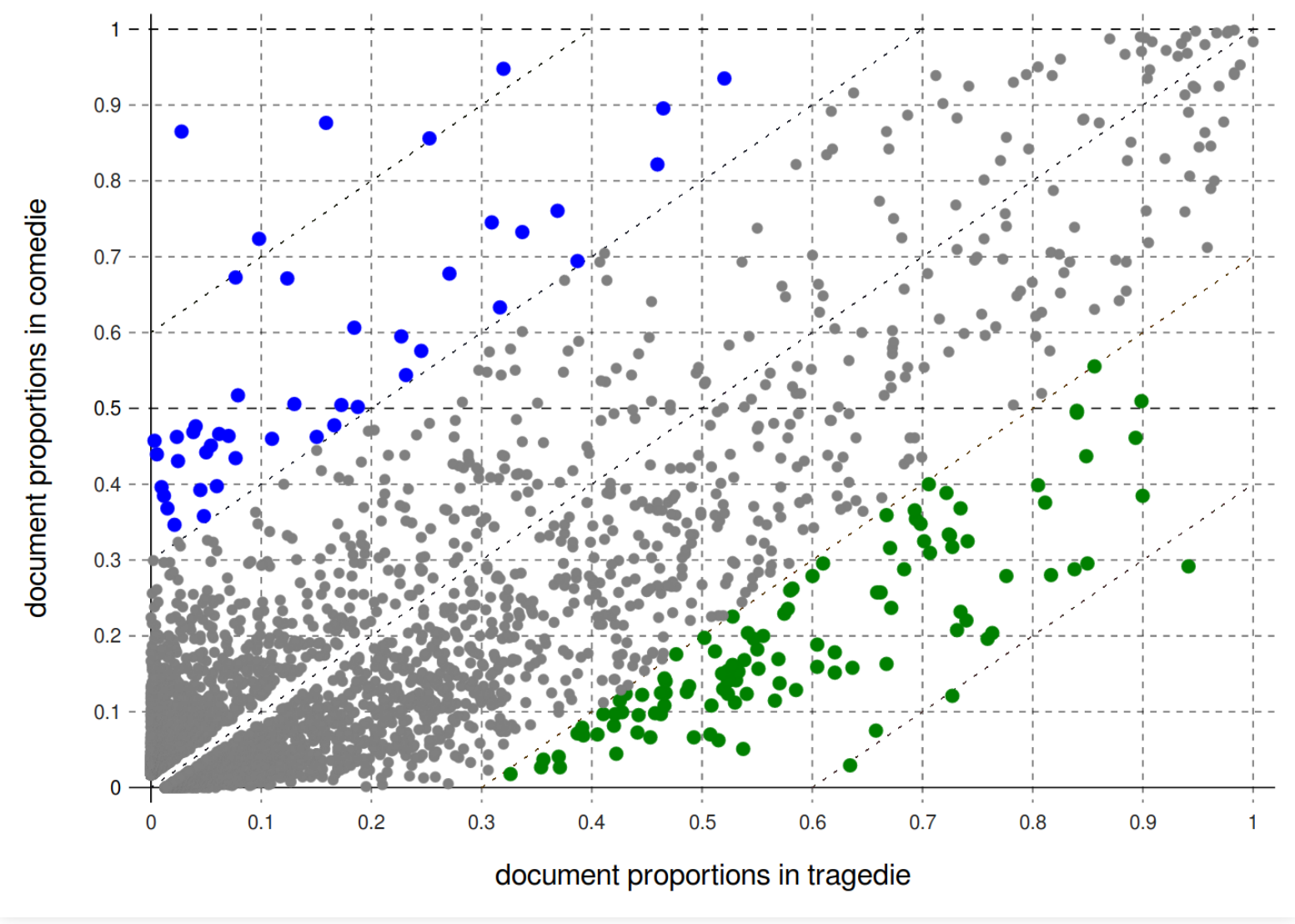
Each dot is one play

PCA on Zeta: Tragedy, comedy, tragicomedy



Each dot is one play

Segment proportions and Zeta



Each dot is one word

3. Variants and evaluation

Relevant parameters

- Segment size: m words
- (Sampling method for the segments)

Possible variants of Zeta

- use relative frequencies instead of segment proportions
- use division instead of subtraction
- use log-transformed values instead of untransformed values

Overview of variants

	segment proportions		relative frequencies	
	normal values	**log2 values**	**normal values**	**log2 values**
subtraction	sp0	sp2	sr0	sr2
division	dp0	dp2	dr0	dr2

sp0 = Burrows Zeta, sp2 = log2-Zeta

Possible desired effects

- improve distinctiveness
- maintain interpretability

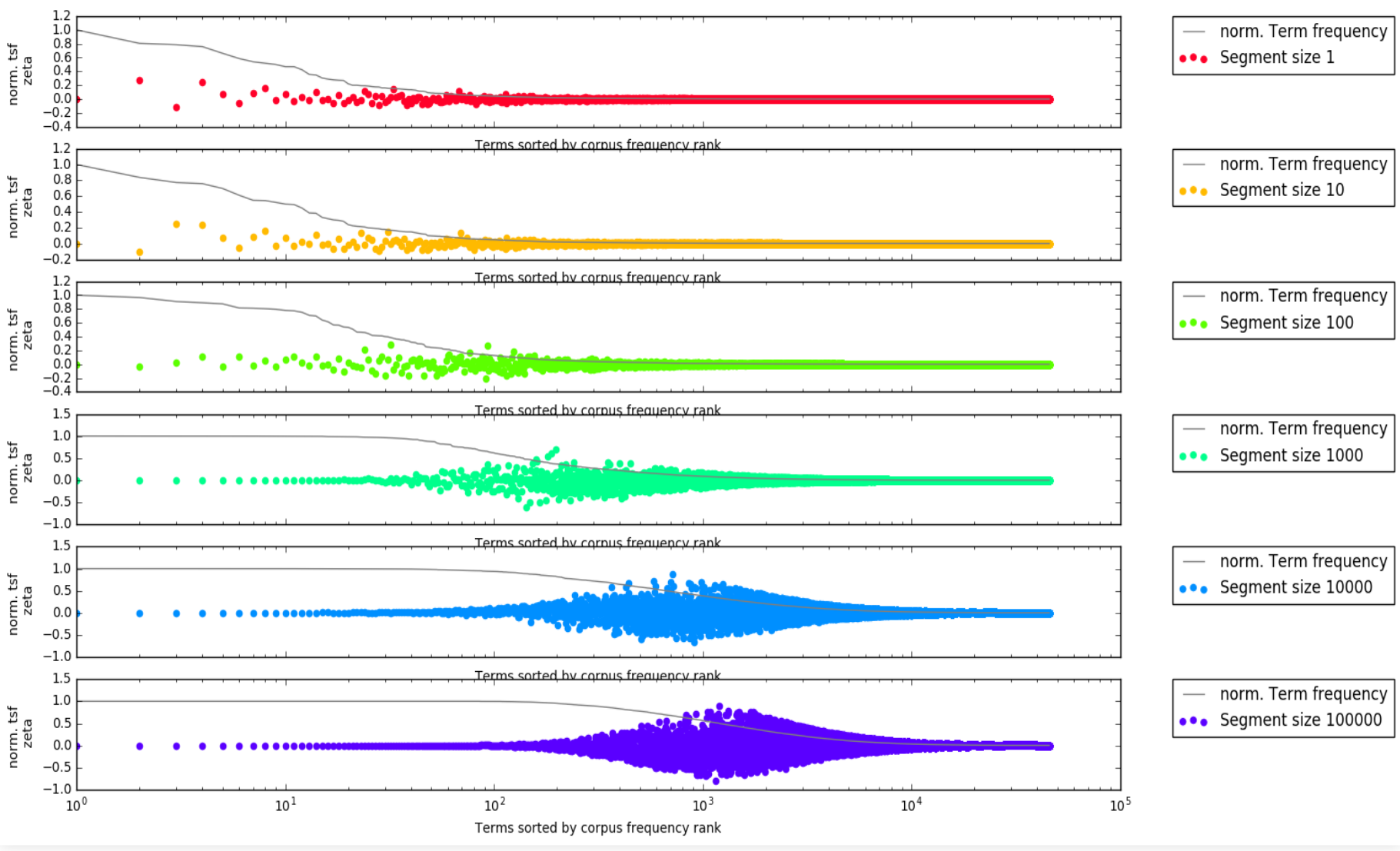
Text collection used

- Today: results from a collection of Spanish novels
- Date of publication: 1880-1940
- 24 novels from Spain, 24 novels from Latin America
- Source: CLiGS textbox, github.com/cligs/textbox (Ulrike Henny-Krahmer and José Calvo Tello)

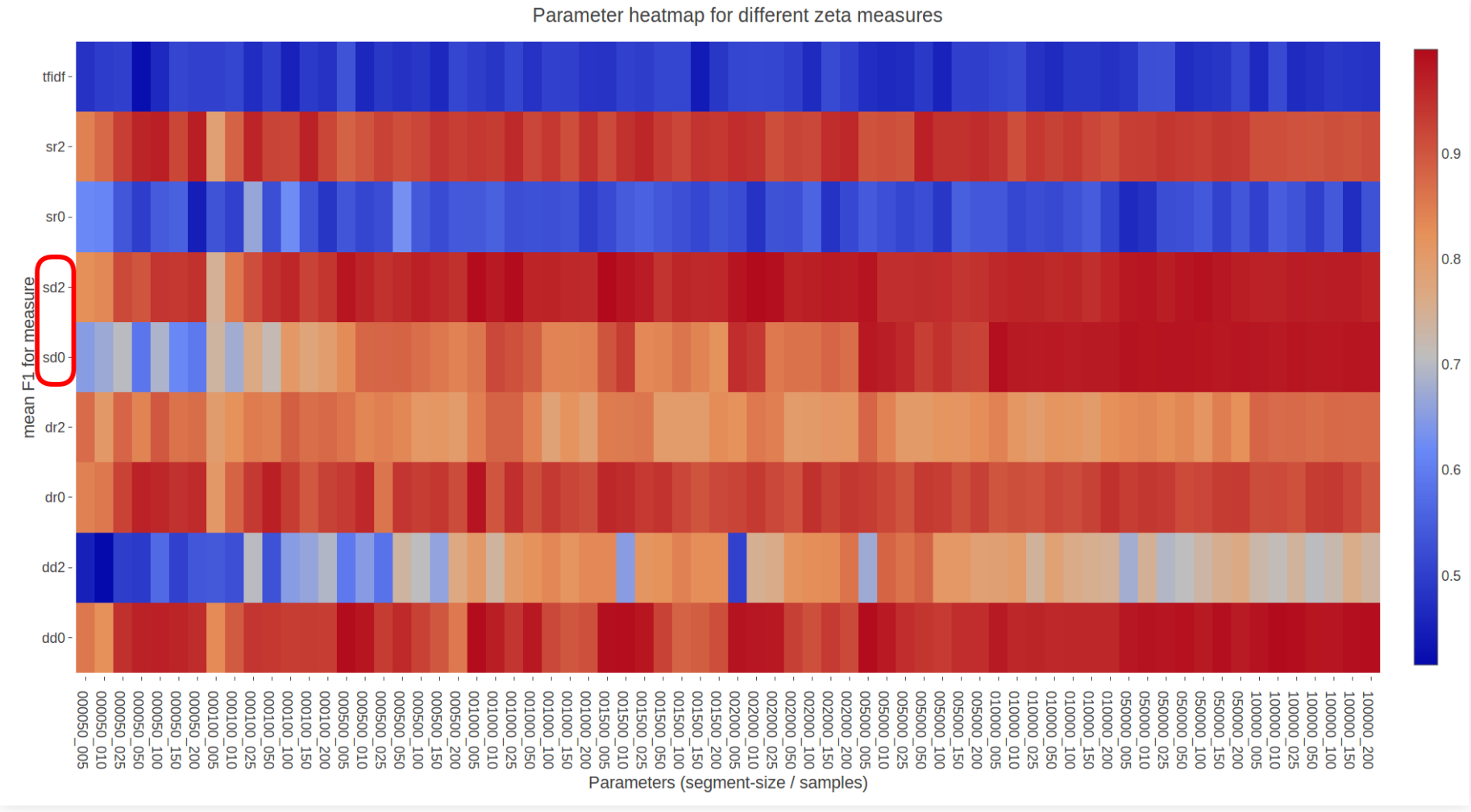
Methods

- Exploratory
  - plot Zeta data, varying parameters and variants
  - aim: better understanding
- Performance testing
  - classification task with varying parameters and variants
  - aim: find out whether some variants perform better

Exploratory: Zeta and segment size



Performance: Classification task



Zeta variants (rows) and parameters (columns)

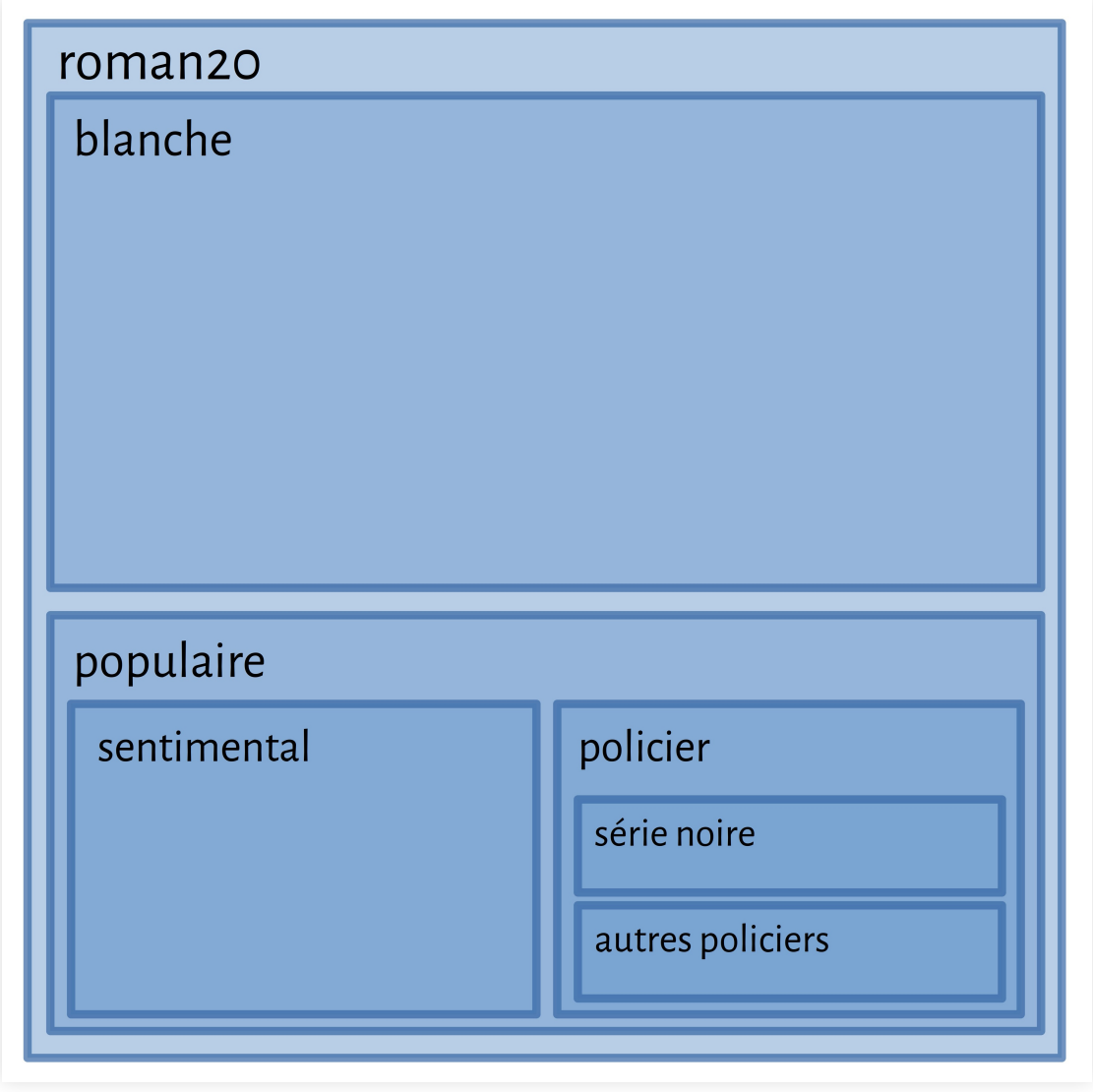
Task details: linear SVM classifier using 40 most distinctive words, three-fold cross-validation; tf-idf Baseline 0.49

4. Application: French Contemporary Novel

Data

- Corpus of French Contemporary Novel, 1950-2010 (in progress): currently 800 novels; target: ca. 3000 novels
- Today: preliminary results from a collection of 180 novels

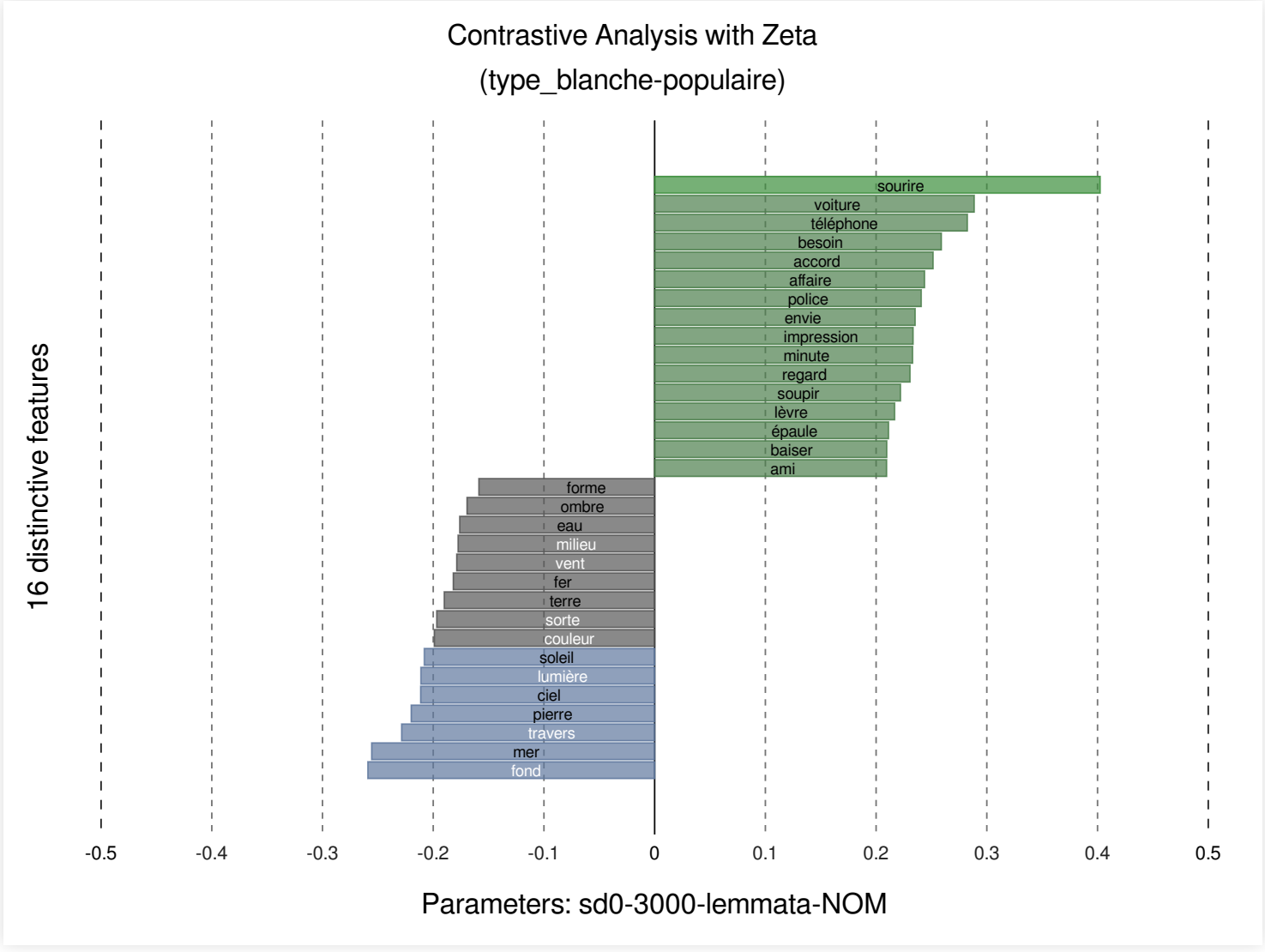
Overview of the corpus



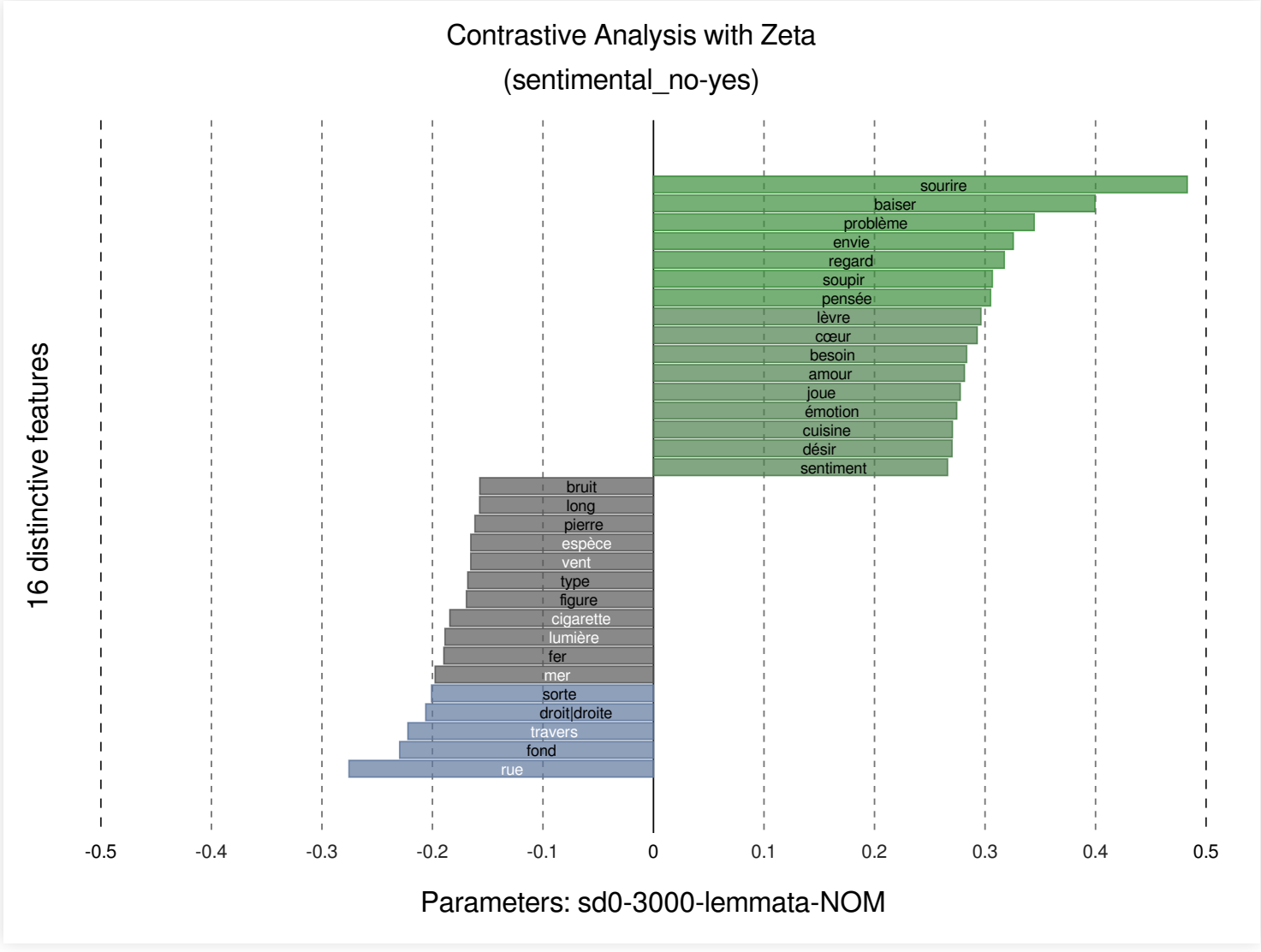
Research questions

- What are words characteristic of low-brow novels (populaire) when compared to high-brow novels (blanche)?
- What are words characteristic of crime fiction (policier) when compared to non-crime fiction (sentimental)?
- What are words characteristic of "série noire" novels when compared to other crime fiction?

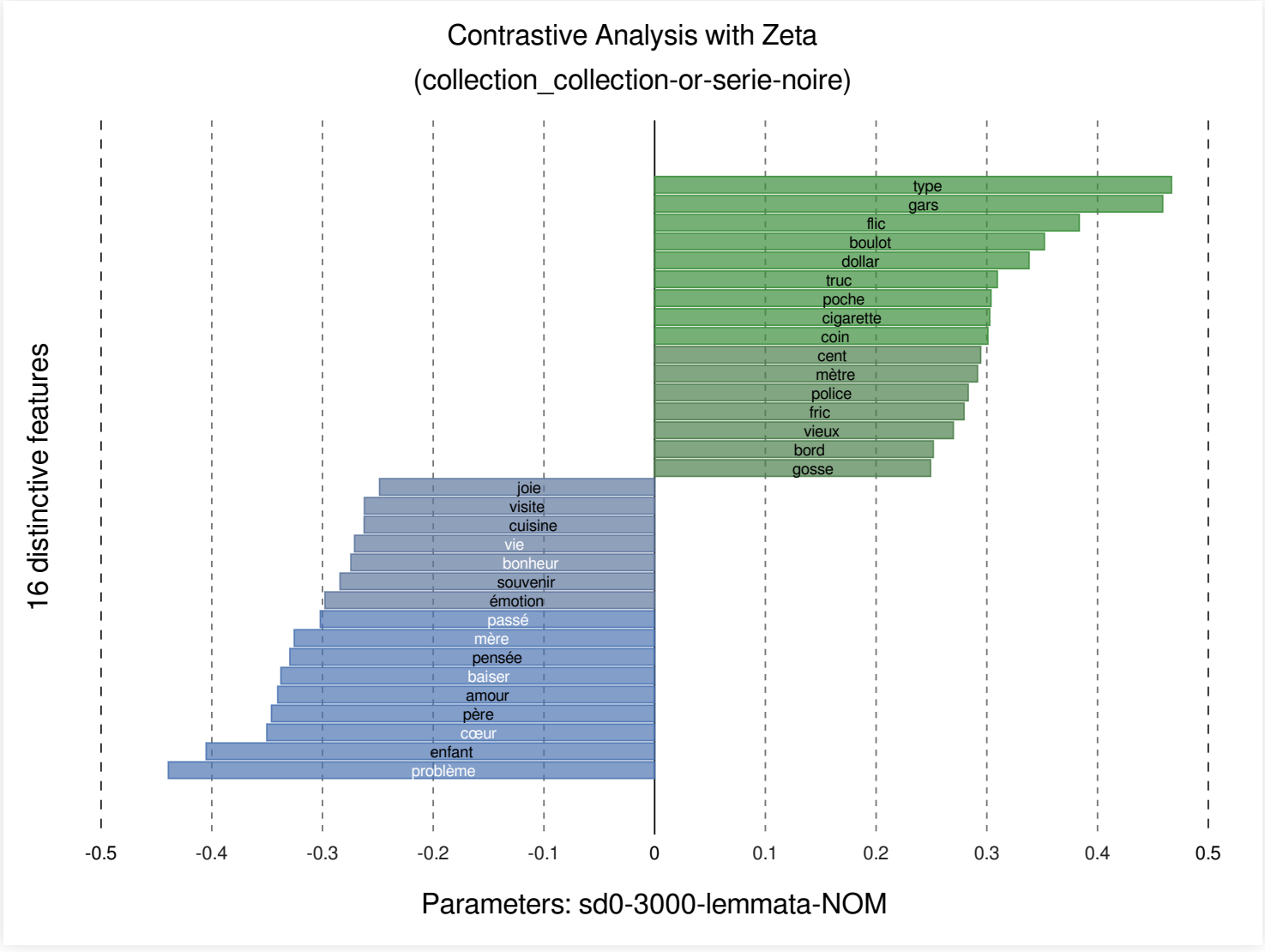
blanche vs. populaire



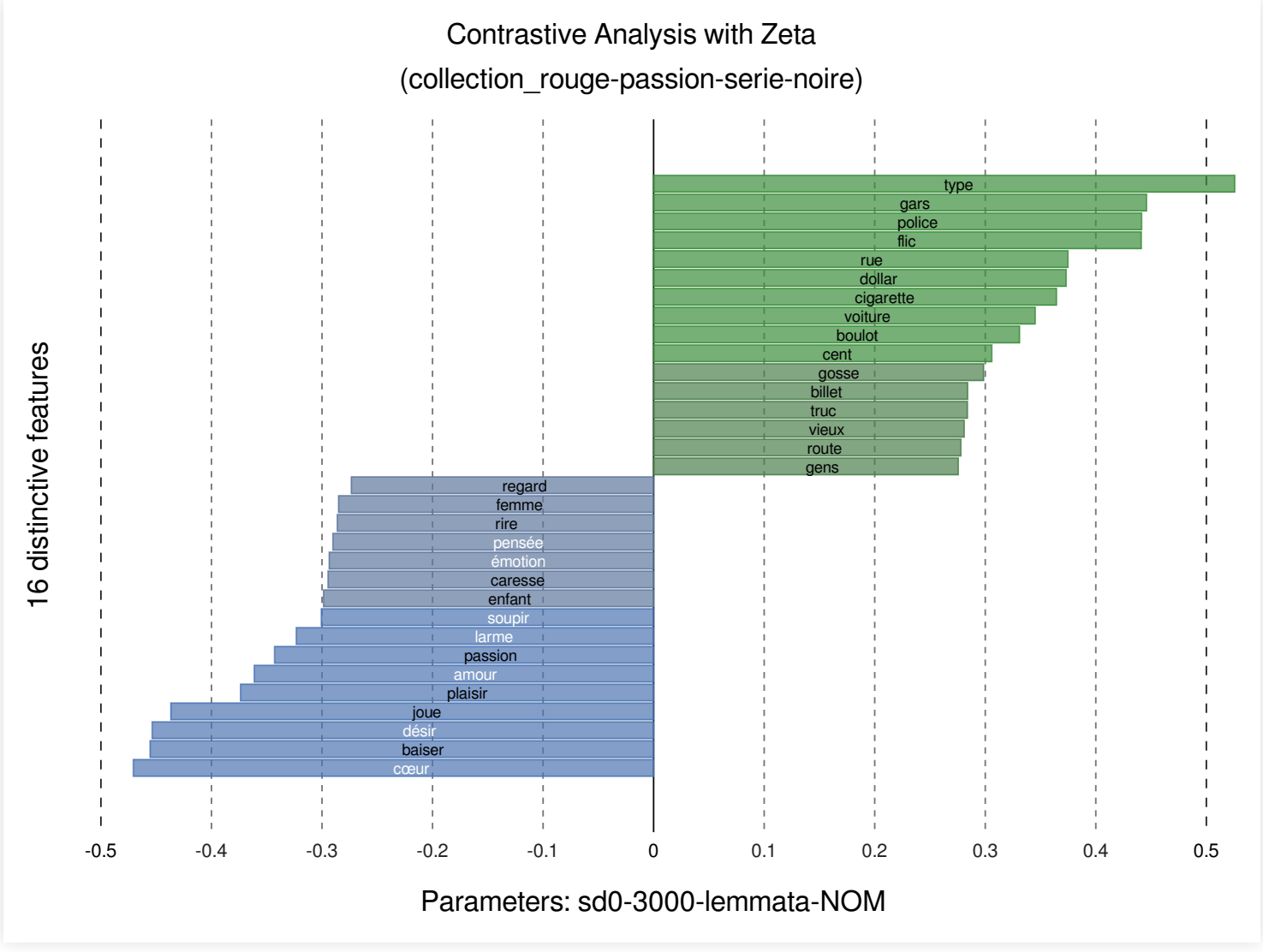
sentimental vs. non-sentimental



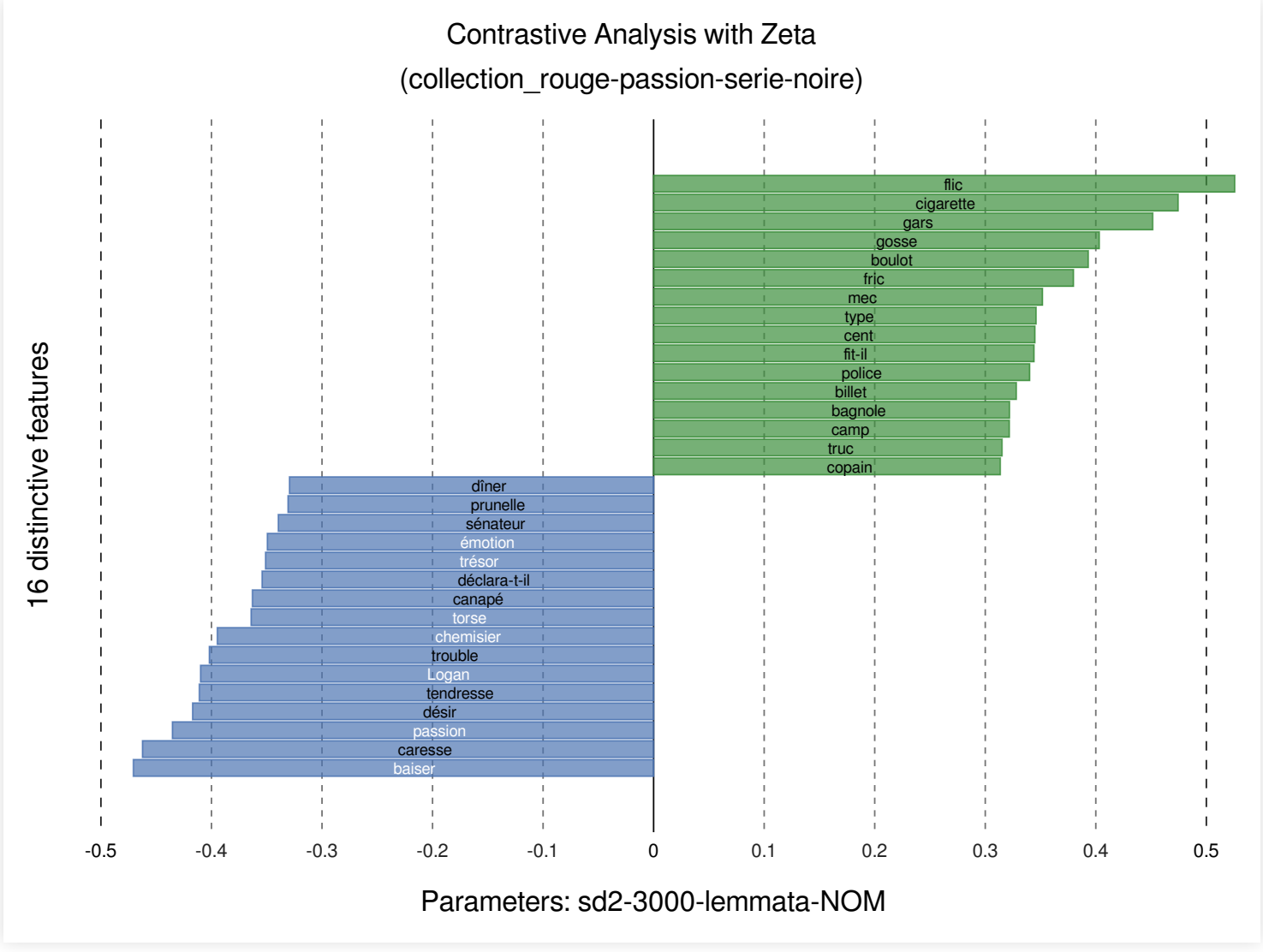
policier vs. sentimental (coll. or)



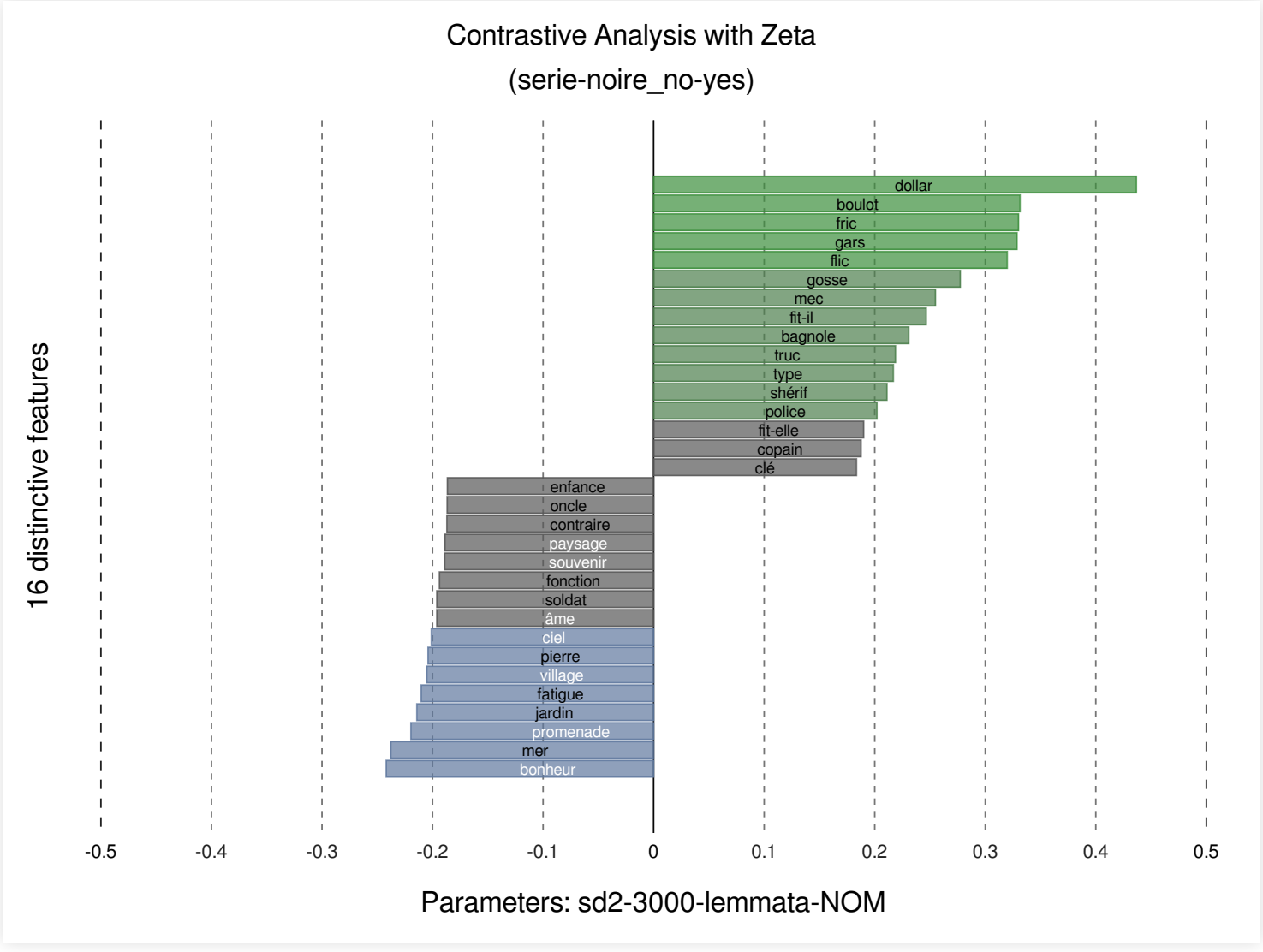
policier vs. sentimental (passion rouge, sdo)



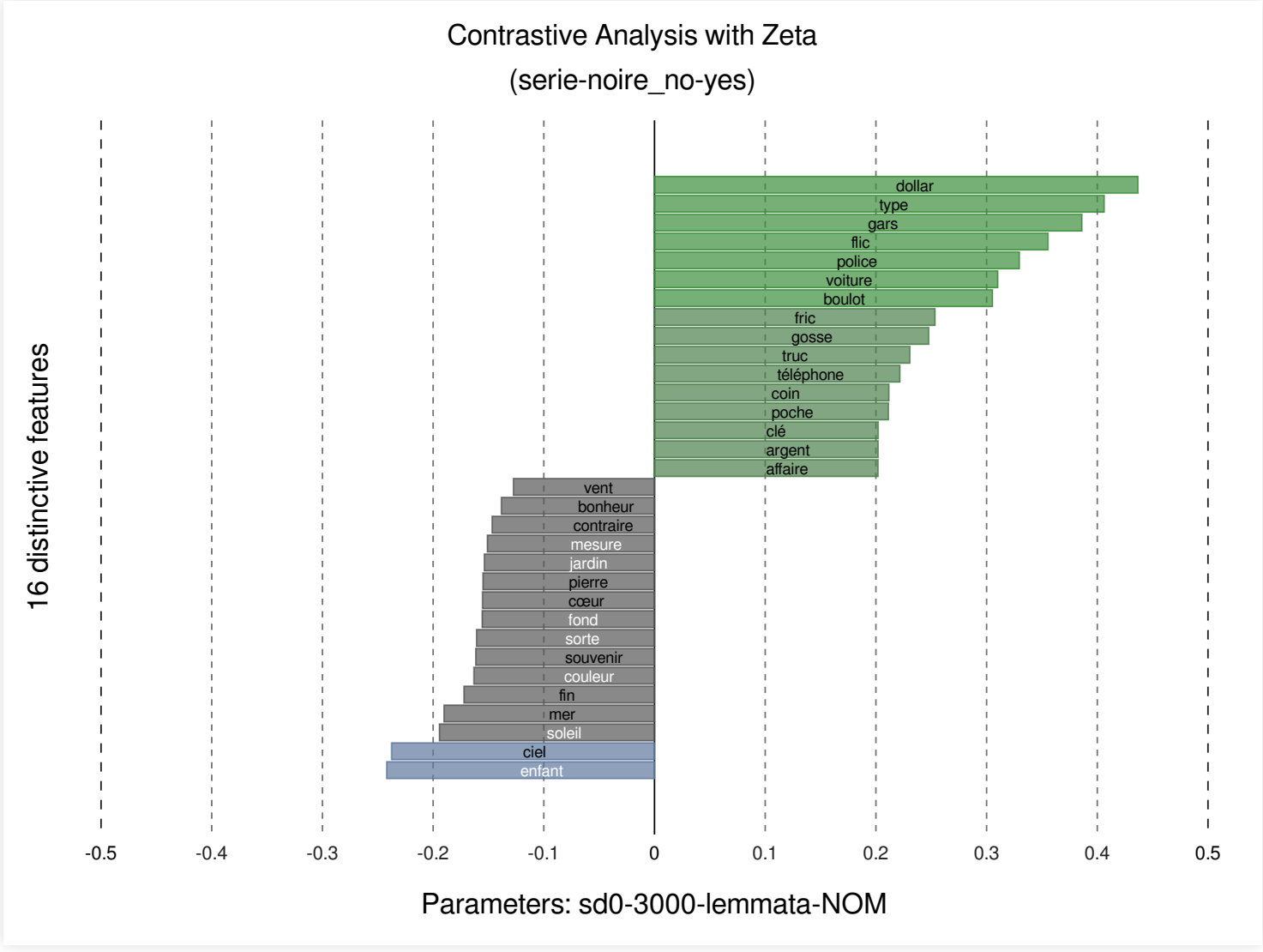
policier vs. sentimental (passion rouge, sd2)



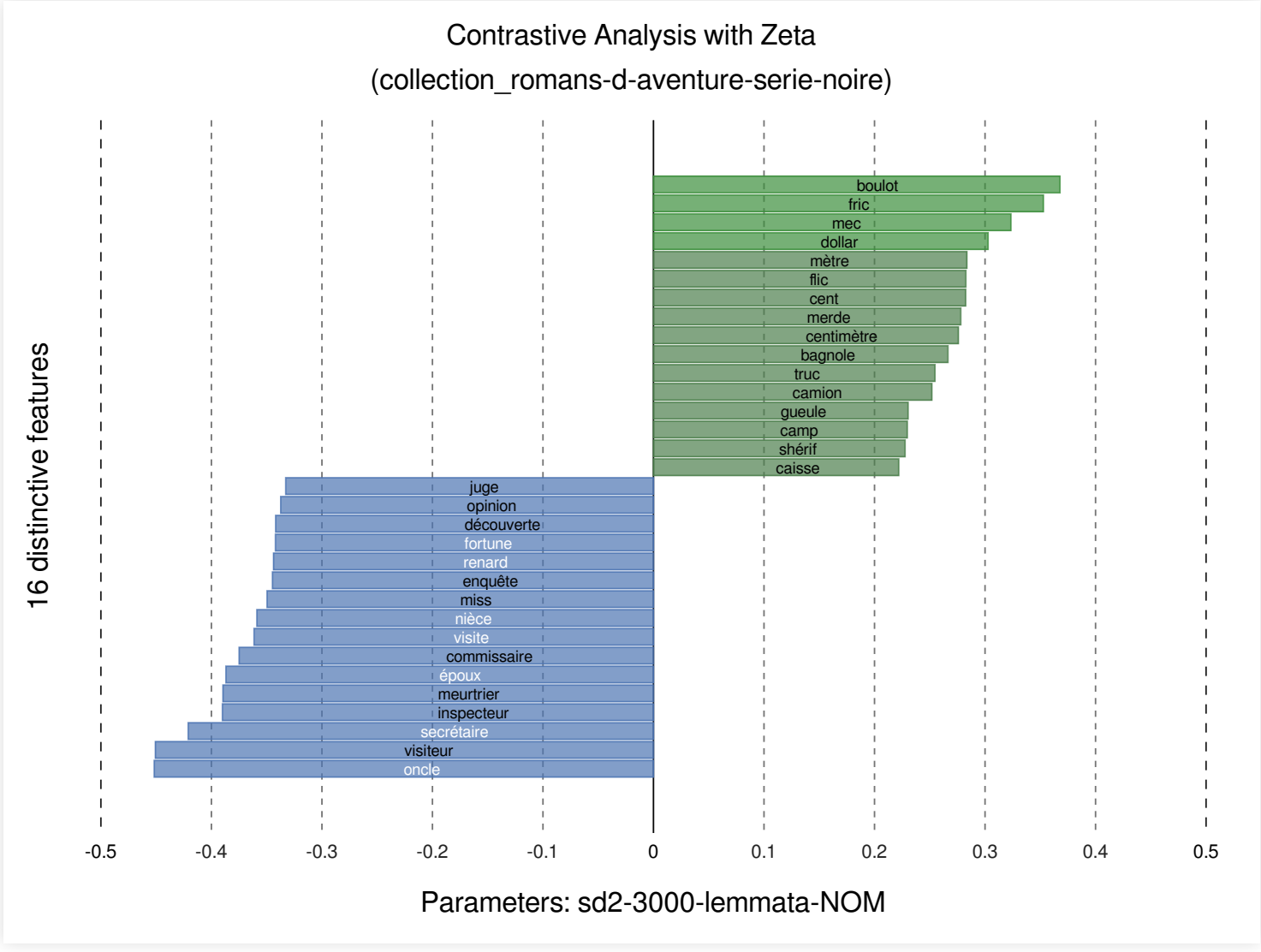
série noire vs. other novels (sd2)



série noire vs. other novels (sdo)

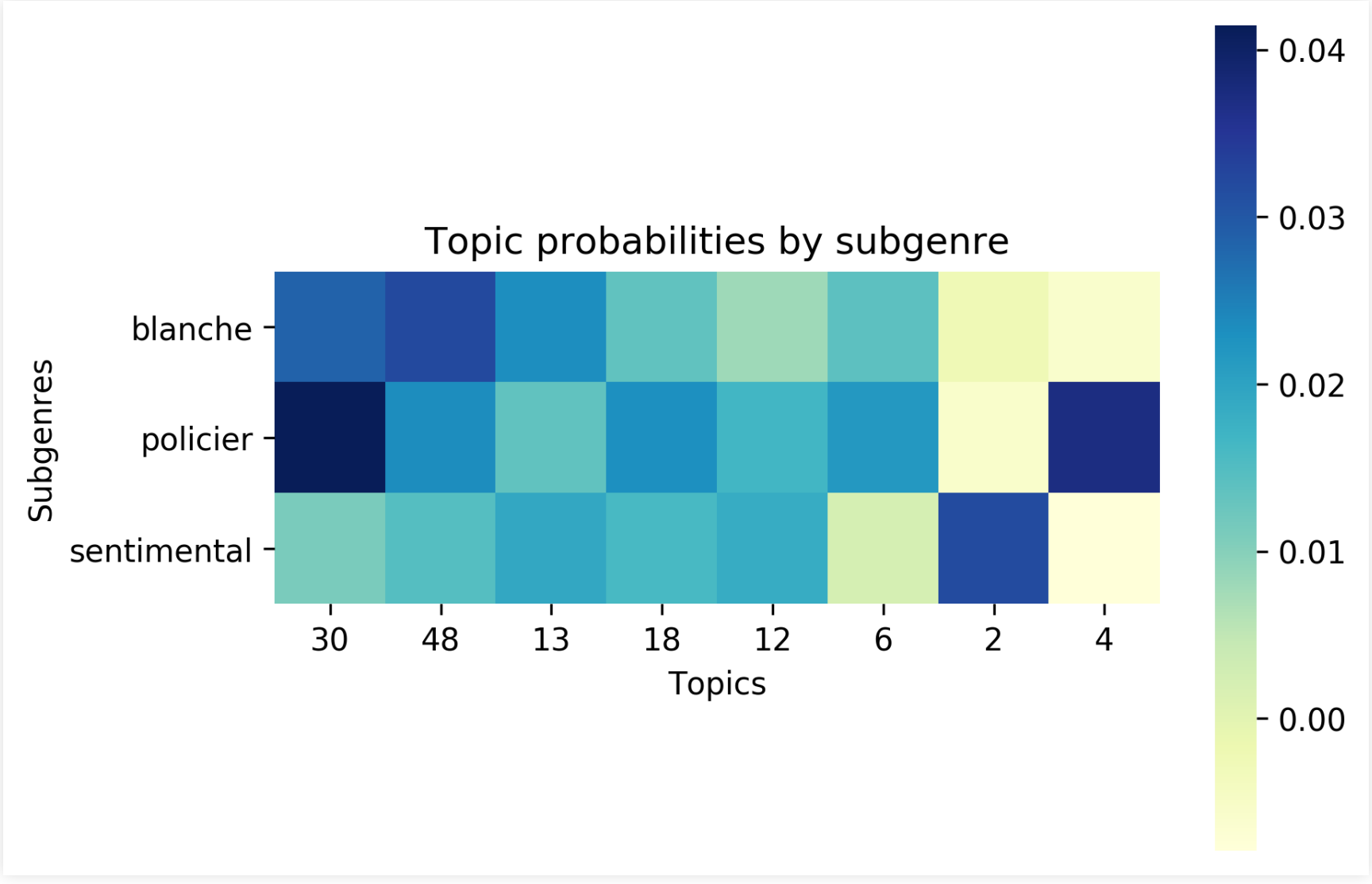


série noire vs. autres policiers



Alternative: Topic Modeling

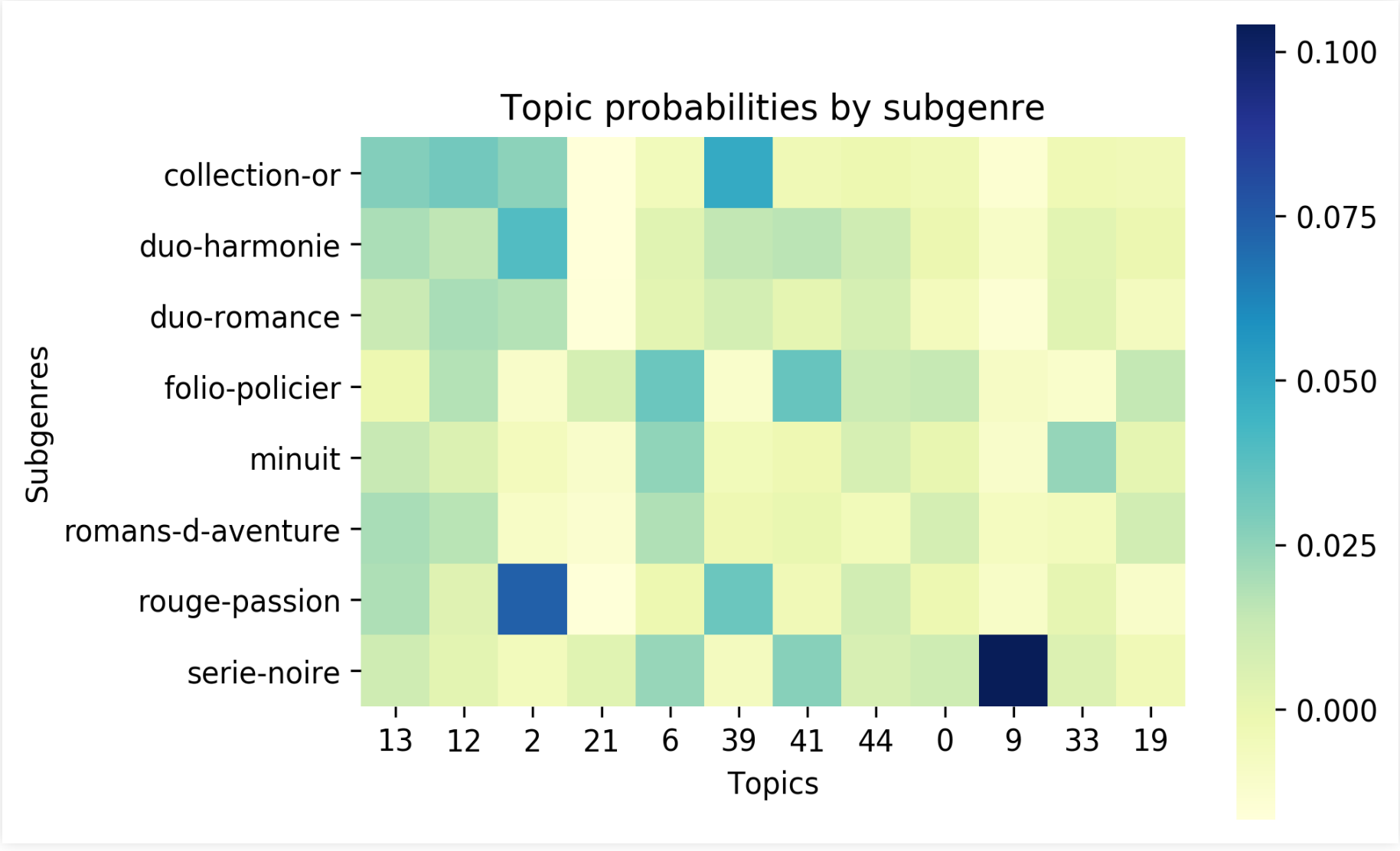
Subgenres


Topic probabilities by subgenre
Subgenres
blanche
policier
sentimental
30
48
13
18
12
6
2
4
Topics
0.04
0.03
0.02
0.01
0.00

Collections: Topics

topic 30	topic 4	topic 2
main œil tête	police affaire	main bras œil
coup bras pied	inspecteur	femme corps
foi jambe fois	commissaire mort	lèvre amour
air visage	policier homme	désir visage
homme corps	enquête bureau	baiser tête cœur
sang bouche	femme crime	cheveu lit nuit
chose moment	question heure nom	bouche doigt
voix temps	chose victime	épaule joue
épaule	meurtre coup ami	peau sein
	personne	

Collections

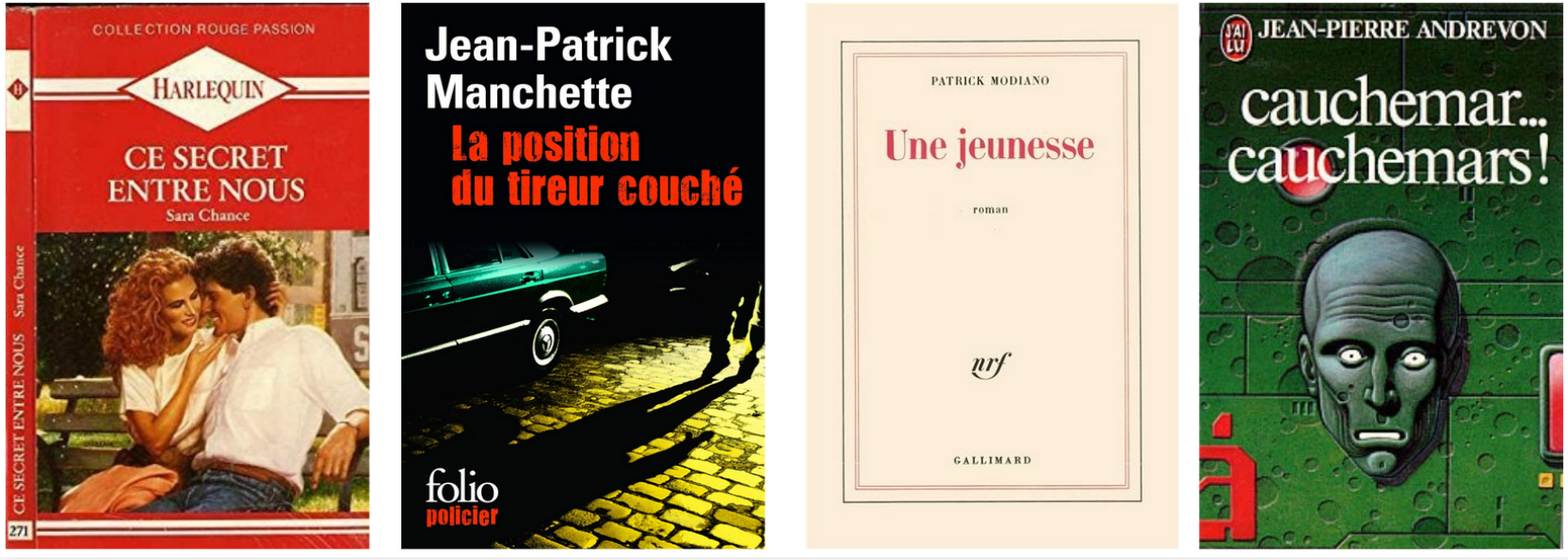


Collections: Topics

topic 9	topic 39	topic 2
dollar argent	père enfant mère	main bras œil
type legs coup	fille frère parent	femme corps lèvre
billet œil air	famille parents	amour désir
tête foi chose	maison maman fil	visage baiser tête
fois banque fille	garçon fils sœur	cœur cheveu lit
affaire gars	femme vie année	nuit bouche doigt
besoin flic	papa grand-mère	épaule joue peau
boulot shérif	école	sein

5. Current Work

A new benchmark corpus

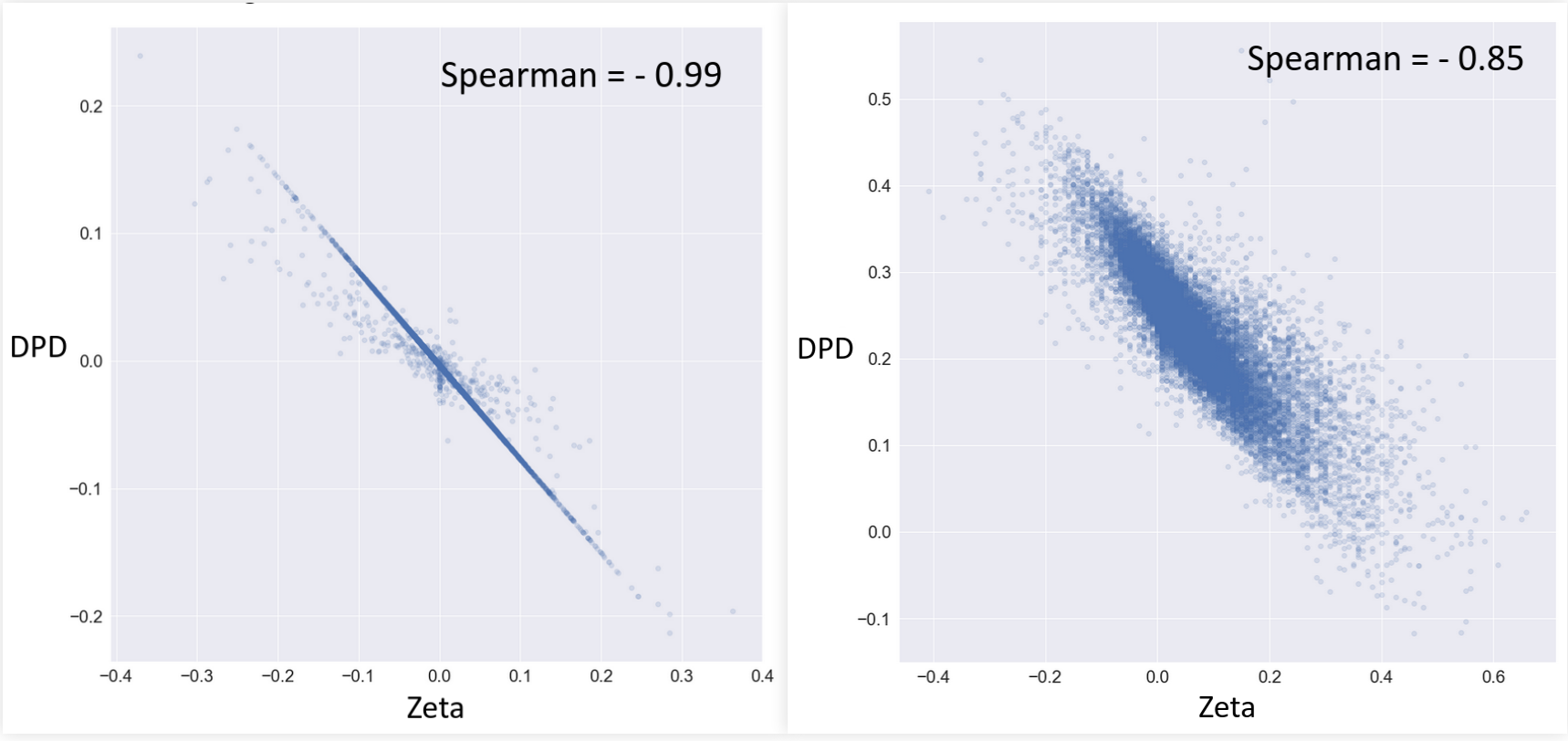


- Contemporary French Novel (1950-2000)
- sentimental, crime, science-fiction, "blanche"
- Currently around 400 / 1200 novels
- To be published in "derived formats"

Exploration of dispersion

- Comparison of dispersion-based measures
- Basis: Zeta (subtraction of document proportions)
- Alternative: 'DPD' (Deviation of proportions distinctiveness; Gries 2008)
  - Gries' DP: a measure of dispersion
  - subtraction of DP scores from each other

Results of the comparison (statistically)



- DPD vs. Zeta scores
- With segmentation (5000 words, left) or with entire novels (right)

Results of the comparison (word-lists)

	Zeta (5000)	Translation
1	humain	human
2	cerveau	brain
3	planète	planet
4	atteindre	achieve; reach
5	centre	center
6	nombre	number
7	système	system
8	émettre	emit
9	univers	universe
10	écran	screen

Translation	DPD (5000)
brain	cerveau
planet	planète
human	humain
center	centre
number	nombre
system	système
emit	émettre
universe	univers
screen	écran
achieve; reach	atteindre

	Zeta (novel)	Translation
1	orbite	orbit
2	civilisation	civilisation
3	terrestre	earthly
4	ordinateur	computer
5	électronique	electronic
6	robot	robot
7	magnétique	magnetic
8	humanité	humanity
9	concept	concept
10	nucléaire	nuclear

X

Translation	DPD (novel)
partial	partiel
chemical	chimique
functioning	fonctionnement
broadcasting	diffusion
diameter	diamètre
hypnotic	hypnotique
radiation	radiation
criterion	critère
govern	régir
vertebral	vertébral

- science-fiction vs. rest
- differences in specificity

Conclusion

Results: Measures

- We have gained a more precise understanding of Zeta
- relation between segment proportions and Zeta (glass ceiling effect)
- a motivated variant of Zeta: log2-Zeta (better performance and robustness)
- a new dispersion-based measure: DPD

Results: Application

- New insights into relationship tragedy / comedy / tragicomedy
- Subtle differences between variants, with dependence on segment length

Next steps

- Tackle the issue of "interpretability": what is it, how can we operationalize it? Is there a trade-off?
- Systematic evaluation of Zeta and (around a dozen) similar measures

Many thanks! | References

- Burrows, John F. (2007). "All the way through: testing for authorship in different frequency strata". *Literary and Linguistic Computing*, 22(1): 27-48.
- Egbert, J., and Biber, D. (2019). “Incorporating text dispersion into keyword analysis”. Corpora, 14(1), 77-104.
- Gries, Stefan, ‘A New Approach to (Key) Keywords Analysis: Using Frequency, and Now Also Dispersion’, Research in Corpus Linguistics, 9 (2021), 1–33 https://doi.org/10.32714/ricl.09.02.02
- Hoover, David L. “Teasing out Authorship and Style with T-Tests and Zeta.” In *Digital Humanities Conference*. London, 2010. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-658.html.
- Lijffijt, Jefrey et al. “Significance Testing of Word Frequencies in Corpora.” *Digital Scholarship in the Humanities* 31, no. 2 (2014): 374–97. doi:10.1093/llc/fqu064.
- Paquot, M., and Bestgen, Y. (2009). “Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction”. In Corpora: Pragmatics and discourse. Brill Rodopi, 247-269.
- Rayson, Paul, and R. Garside. “Comparing Corpora Using Frequency Profiling.” In *Proceedings of the Workshop on Comparing Corpora*, 1–6. Hong Kong: ACM, 2000.
- Schöch, Christof. „Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie“, in: *Quantitative Ansätze in den Literatur- und Geisteswissenschaften*, hg. Toni Bernhard et al. Berlin: de Gruyter, 2018.
- Schöch, C., Schlör, D., Zehe, A., Gebhard, H., Becker, M. & Hotho, A. (2018). “Burrows’ Zeta: Exploring and Evaluating Variants and Parameters”. Digital Humanities Conference: Book of Abstracts: 274-277.

With special thanks to pygal and reveal.js

Thank you!

Slides: christofs.github.io/pyzeta_en/
Code: http://github.com/cligs/pyzeta

Christof Schöch, 2019-2021
