Universität Trier

Deviation of Proportions as the Basis for a Keyness Measure

Keli Du, Julia Dudar, Cora Rok, Christof Schöch
('Zeta and Company' | https://zeta-project.eu | Trier University)

43rd Annual Conference of the German Linguistic Society (DGfS)
25 - 26 February 2021, University of Freiburg

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation

Kompetenzzentrum
für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften

Computational LiteraryStudies

Content

1. Introduction: The "Zeta and Company" Project
2. Our idea: Burrows’ Zeta + Gries’ DP → DP-Distinctiveness (DPD)
3. Application and results
    a. Statistics
    b. Interpretation of the word lists
4. Conclusion

(1) Introduction:

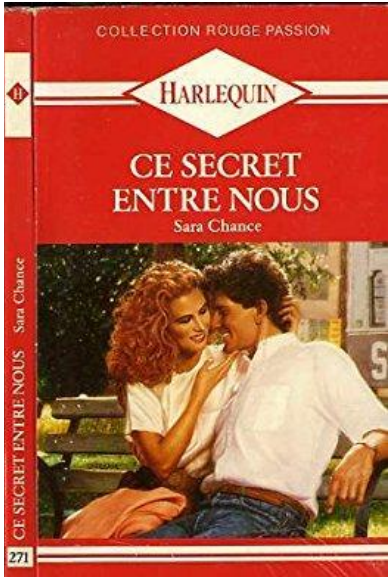
The "Zeta and Company" Project

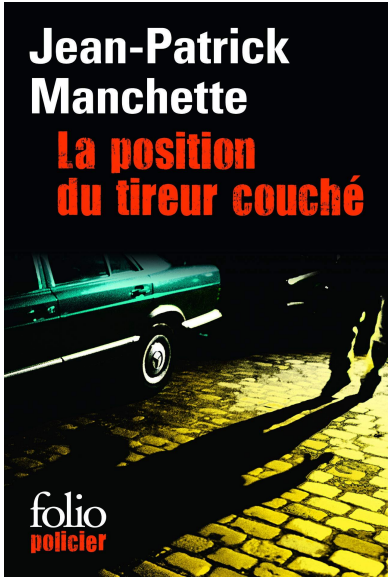
Zeta and Company

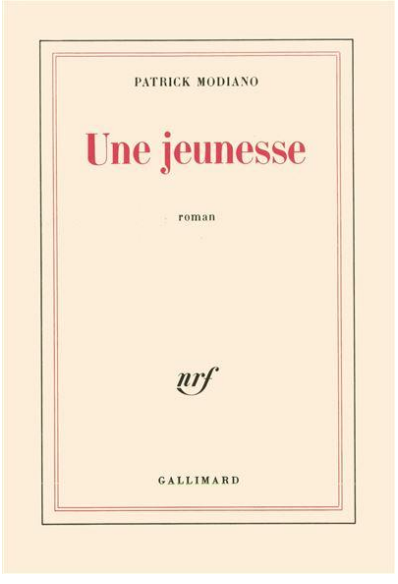
- *Zeta and Company: Measures of Distinctiveness for Computational Literary Studies* (2020-2023); see: https://zeta-project.eu
- Part of the DFG Priority Programme *Computational Literary Studies* (SPP 2207); see: https://dfg-spp-cls.github.io/
- Key objective: Model, implement, evaluate, and use of 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

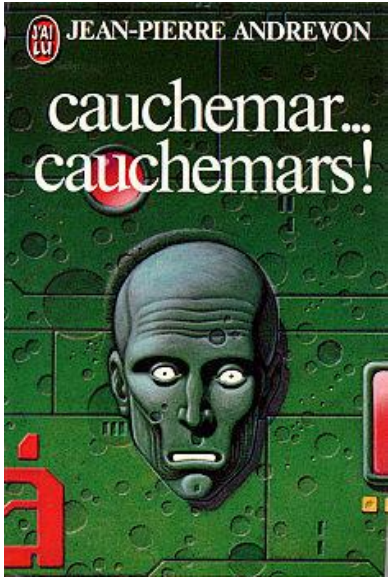
Corpus

- French novels 1980-1989
- 4 groups: sentimental, crime, scifi, high-brow
- Size: 4 x 40 = 160 novels (extension in progress)









(2) Our idea:

Burrows' Zeta + Gries' DP

→ DP-Distinctiveness (DPD)

Dispersion and the research idea

- Keyness measures based only on frequency can be misleading;
- Distribution of words must be considered as well (Egbert & Biber 2019);
- Dispersion: the degree of even distribution of a feature (word, POS, lemma)
- Gries (2008): overview of dispersion measures + Deviation of Proportions (DP).
- Our idea: Gries' DP can be used for quantitative comparative analysis of two text groups;

Gries' deviation of proportions (DP)

- x_i = number of words in text *i*
- l = number of words in corpus
- a = the word of interest
- f = frequency of word *a* in corpus
- f_i = frequency of word *a* in text *i*
- s_i = relative size of text *i*: $\frac{x_i}{l}$
- v_i = relative frequency of word *a* in text *i*: $\frac{f_i}{f}$
- $DP = \frac{\sum_{i=1}^{n} |s_i - v_i|}{2}$

Zeta as dispersion and distinctiveness measure

$docf_i()$ = number of documents in the corpus, where word i occurs at least once;

n = total number of documents in the corpus

$$docp_i(T) = \frac{docf_i(T)}{n(T)} \quad \text{and} \quad docp_i(R) = \frac{docf_i(R)}{n(R)}$$

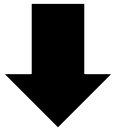
$$Zeta_i = docp_i(T) - docp_i(R)$$

- Zeta: comparison of the dispersion by simply subtracting the two values
- the higher the Zeta-score, the more distinctive the feature

From Zeta to DP-based distinctiveness measure

$$docp_i(T) = \frac{docf_i(T)}{n(T)} \quad \text{and} \quad docp_i(R) = \frac{docf_i(R)}{n(R)}$$

$$Zeta_i = docp_i(T) - docp_i(R)$$



$$DP_i(T) = \frac{\sum_{i=1}^{n} |s_i(T) - v_i(T)|}{2} \quad \text{and} \quad DP_i(R) = \frac{\sum_{i=1}^{n} |s_i(R) - v_i(R)|}{2}$$

$$DPD_i = DP_i(T) - DP_i(R)$$

(3) Application and results

a. Statistics

b. Interpretation of the word lists

Test

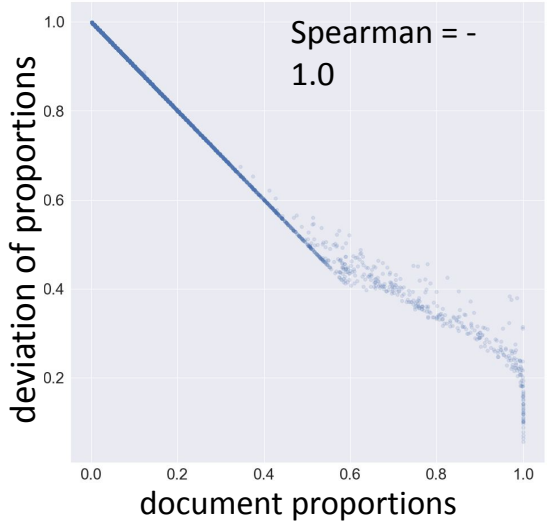
- target corpus (scifi) vs. reference corpus (non-scifi)
  a. 5000-lemmata-segments
  b. 10000-lemmata-segments
  c. novel (> 56000 lemmata in average)
- calculation (Zeta & DPD)
- ranking words by their document proportions, deviation of proportions (DP), Zeta scores, DPD scores
- calculate Spearman's rank correlation between the 4 rankings

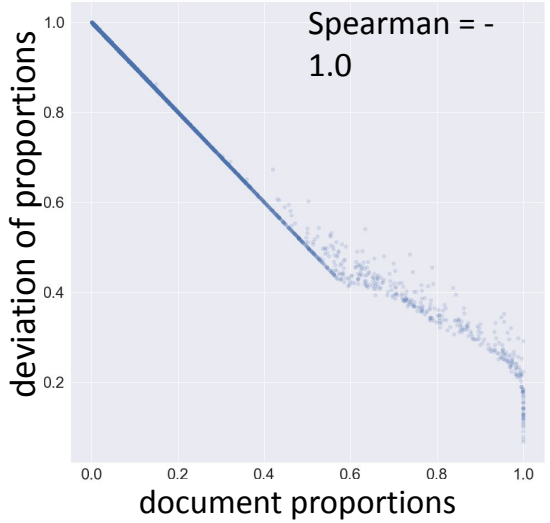
Results / observations

Words' ranking based on:

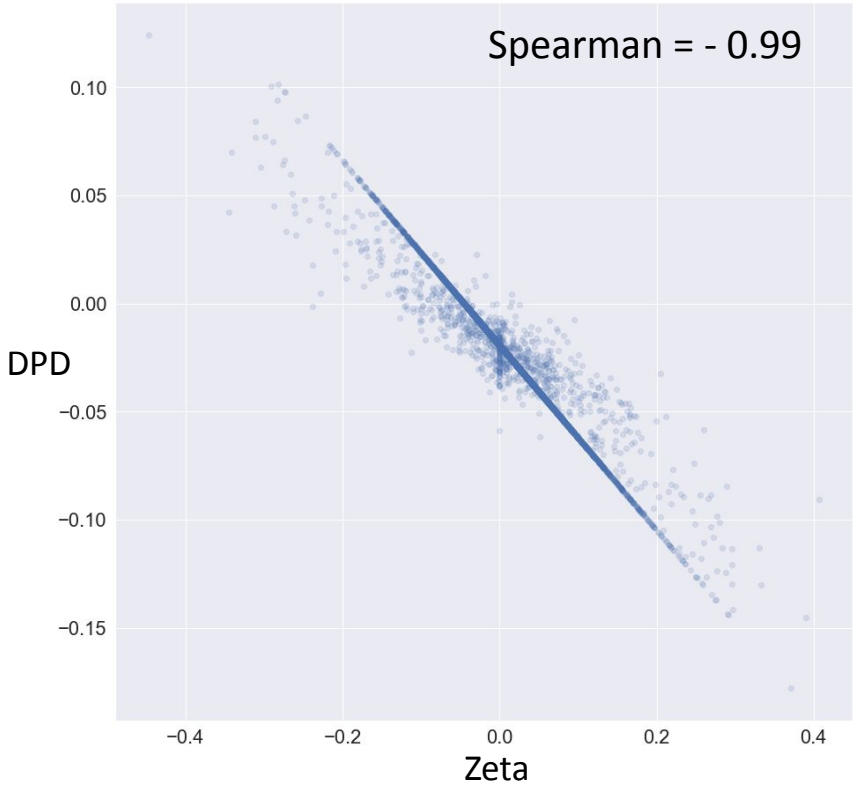
- deviation of proportions (DP) vs. document proportions ⇒ negative correlation
- Zeta vs. DPD ⇒ negative correlation
- The correlation between Zeta ranking and DPD ranking has a tendency to weaken as segment length increases from 5000 lemmata over 10000 lemmata to the novel-level.

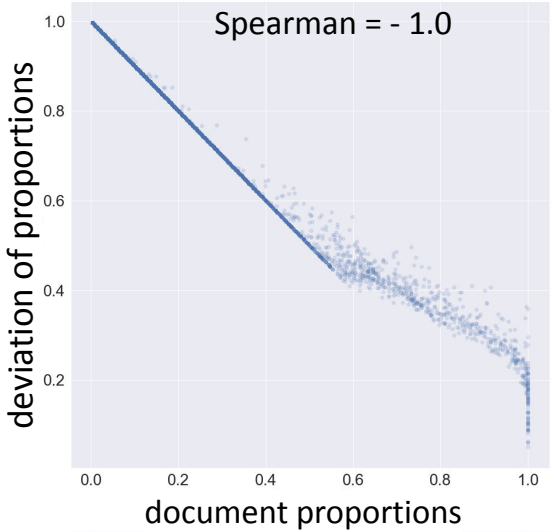
- correlation between:
  - DPD and Zeta (left),
  - DP and document proportions in target corpus (upper right)
  - DP and document proportions in reference corpus (lower right)
- 5000 tokens segments

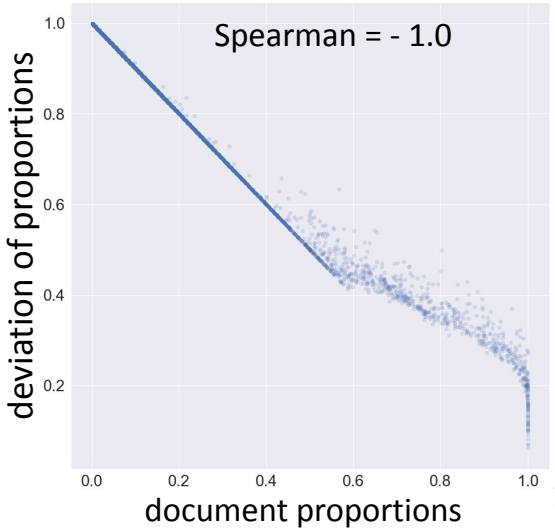




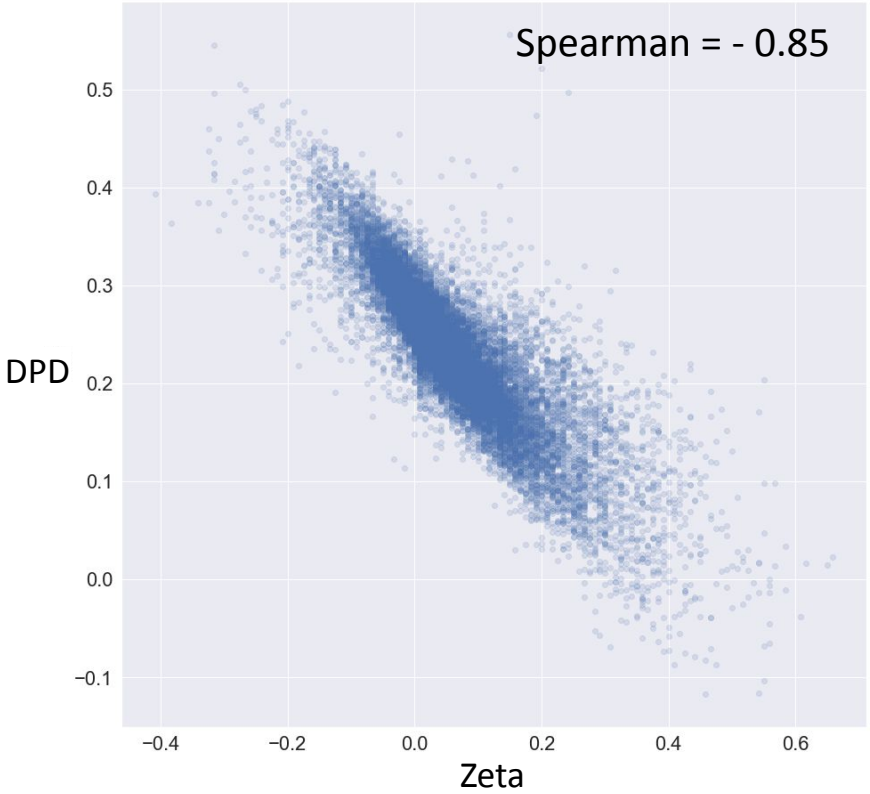
- correlation between:
  - DPD and Zeta (left),
  - DP and document proportions in target corpus (upper right)
  - DP and document proportions in reference corpus (lower right)
- 10000 tokens segments

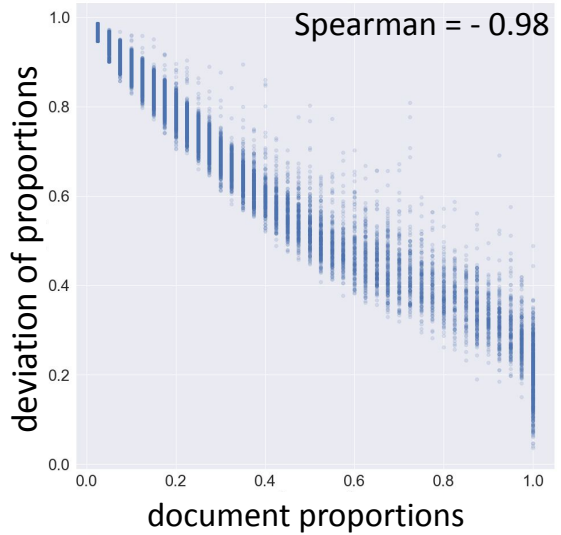






- correlation between:
  - DPD and Zeta (left),
  - DP and document proportions in target corpus (upper right)
  - DP and document proportions in reference corpus (lower right)
- each novel as a segment





(3) Application and results

a. Statistics

b. Interpretation of the word lists

Top 10
Zeta vs. DPD
word list

segment-based
comparison
(5000 tokens)

Science Fiction

→ overlapping top 10 keywords with slightly different ranking

	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

b. Interpretation of the word lists

Science-Fiction (target corpus)
segment-based comparison of Zeta and DPD (5000 tokens)

semantic fields	Zeta | DPD
1) lifeform	human, brain
2) space	planet, universe
3) spatial data	center
4) computation	number
5) technology	system, screen
6) movements	achieve, emit

Top 10 Zeta vs. DPD word list

segment-based comparison (10000 tokens)

Science Fiction

→ less overlapping keywords

DPD ranking

20

11

	Zeta (10000)	Translation
1	humain	human
2	cerveau	brain
3	émettre	emit
4	planète	planet
5	système	system
6	niveau	level
7	univers	universe
8	nombre	number
9	base	base
10	centaine	hundred

Translation	DPD (10000)
emit	émettre
brain	cerveau
hundred	centaine
computer	ordinateur
level	niveau
civilisation	civilisation
electronic	électronique
function	fonctionner
complex	complexe
planet	planète

Zeta ranking

11

18

19

21

28

b. Interpretation of the word lists

Science-Fiction (target corpus)
segment-based comparison of Zeta and DPD (10000 tokens)

→ all keywords can be assigned to the previously established semantic categories:

1) lifeform	Zeta: human, brain	DPD: brain, civilisation
2) space	Zeta: planet, universe	DPD: planet
3) spatial data	Zeta: level, base	DPD: level, complex
4) computation	Zeta: number, hundred	DPD: hundred
5) technology	Zeta: system	DPD: computer, function, electronic
6) movements	Zeta: emit	DPD: emit

Top 10
Zeta vs. DPD
word list

Comparison based on whole novels

Science Fiction

→ top 10 keywords don't match (two words of each list ranked among the top 25)

DPD ranking

14

23

	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

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

Zeta ranking

19

14

b. Interpretation of the word lists

Science-Fiction (target corpus)
Comparison based on whole novels

→ top keywords don't fit into the same semantic categories

	Zeta	DPD
lifeform	humanity, civilisation	
space	orbit, earthly	
computation		partial, diameter
technology	computer, electronic, robot	functioning
movements		govern
physics	magnetic, nuclear	
chemistry		chemical, diffusion
astronomy		radiation
anatomy		vertebral
psychology		hypnotic
theories	concept	criterion

Zeta's and DPD's top keywords of the novel-based comparison differ considerably from the keywords of the top word lists of the segment-based comparison

	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

(4) Conclusion

Impact of segment length

- The shorter the segments,
  → the more similar the word lists of Zeta and DPD
  → the more general the keywords in both of the lists

- The longer the segments,
  → the more divergent the word lists of Zeta and DPD
  → the more specific the keywords, especially in DPD's word list

- The novel-based comparison has shown:
  Zeta: → keywords are more general
  → same semantic fields as keywords of the segment-analyses
  DPD: → word list contains scientific terminology
  → new/more semantic fields

Zeta vs. DPD

- Statistics
  - both are dispersion-based
  - although mathematically defined in different ways
  - they have a very strong correlation (for short segments),
  - → what explains similarities in word lists

- Distinctiveness
  - both equally able to extract meaningful and interpretable words
  - both equally adequate to perform a genre analysis

Future Work

- compare Zeta, DPD with other distinctive measures / keyness measures and their variants

- systematically investigate the influence of segment length on distinctive measures

Thank you for your attention!

References

Egbert, J., & Biber, D. (2019). “Incorporating text dispersion into keyword analysis”. Corpora, 14(1), 77-104.

Gries, S (2008). “Dispersions and adjusted frequencies in corpora”. International Journal of Corpus Linguistics, Volume 13(4): 403–437. DOI: https://doi.org/10.1075/ijcl.13.4.02gri

Lijffijt, J., Nevalainen, T., Säily, T., Papapetrou, P., Puolamäki, K. and Mannila, H. (2014). “Significance testing of word frequencies in corpora”. Digital Scholarship in the Humanities, 31(2): 374–97. doi:10.1093/llc/fqu064.

Lyne, A. A. (1985). “Dispersion”. The Vocabulary of French Business Correspondence. Paris: Slatkine / Champion, 101–24.

Newman, M. L., Groom, C. J., Handelman, L. D., and Pennebaker, J. W. (2008). “Gender differences in language use: an analysis of 14,000 text samples”. Discourse Processes, 45: 211–36.

Oakes, M. P. and Farrow, M. (2007). “Use of the chi-squared test to examine vocabulary differences in English-language corpora representing seven different countries”. Literary and Linguistic Computing, 22(1): 85–100.

Paquot, M., & 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, P., Leech, G., and Hodges, M. (1997). “Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus”. International Journal of Corpus Linguistics, 2(1): 133–52.

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.