Burrows’ Zeta: Exploring and Evaluating Variants and Parameters

Christof Schöch (schoech@uni-trier.de), University of Trier, Germany and Daniel Schlör (daniel.schloer@informatik.uni-wuerzburg.de), University of Würzburg, Germany and Albin Zehe (zehe@informatik.uni-wuerzburg.de), University of Würzburg, Germany and Henning Gebhard (s2hegebh@uni-trier.de), University of Trier, Germany and Martin Becker (becker@informatik.uni-wuerzburg.de), University of Würzburg, Germany and Andreas Hotho (hotho@informatik.uni-wuerzburg.de), University of Würzburg, Germany

1. Introduction

The research presented here concerns methodological issues surrounding Zeta, a measure of distinctiveness or keyness initially proposed by John Burrows (2007). Such measures are used to identify features (e.g. words) that are characteristic of one group of texts in comparison to another (Scott 1997), a fundamental task that many standard tools support, e.g. WordCruncher (Scott 1997), AntConc (Anthony 2005), TXM (Heiden et al. 2012) or stylo (Eder et al. 2016). Widely used methods include the log-likelihood ratio (where observed frequencies are compared to expected frequencies; see Rayson and Garside 2000) and Welch's t-test (where two frequency distributions are compared; see Ruxton 2006). Zeta, by contrast, is based on a comparison of the degrees of dispersion of features (see Lyne 1985, Gries 2008). Zeta appeals to the Digital Literary Studies community because it is mathematically simple and has a built-in preference for highly interpretable content words. Indeed, Zeta has been successfully applied to various issues in literary history (e.g. Craig & Kinney 2009, Hoover 2010, Schöch 2018). However, its statistical properties are not well understood, as important work on evaluating measures of distinctiveness (Kilgariff 2004, Lijfijt et al. 2014) has not included Zeta. Therefore, we submit two key aspects of Zeta to exploration and evaluation: (a) variations in the way Zeta is calculated and (b) variations of key parameters. We gain a more precise understanding of how Zeta works and propose a new variant, “log2-Zeta”, that shows more stable behavior for different parameters than Burrows’ Zeta.

2. Deriving Zeta variants and key parameters

Zeta is calculated by comparing two groups of texts (G1 and G2). From each text in each group, a sample of n segments of fixed size with m word tokens is taken. For each term (t) in the vocabulary (e.g., consisting of lemmatized words), the segment proportions (sp) in each group are calculated, i.e. the proportion of segments in which this term appears at least once (binarization). Zeta of t results from subtracting the two segment proportions:

$$\mathrm{zeta}_t = \mathrm{sp}_t(G_1) - \mathrm{sp}_t(G_2)$$

From this formalization, we can derive several variants of Zeta: applying division instead of subtraction; using relative frequencies (rf) instead of segment proportions (sp); and applying a log-transformation to the values rf and sp instead of using them directly. This results in eight variants of Zeta (Table 1).

	segment proportions		relative frequencies	
	normal	log2	normal	log2
subtraction	sd0	ds2	sr0	sr2
division	dd0	dd2	dr0	dr2

Table 1: The eight variants of Zeta with their labels; “sd0” corresponds to Burrows’ Zeta.

The formalization also points to two major parameters of Zeta: segment sampling strategy (using all possible consecutive segments, or sampling n segments per text to overcome text length imbalances) and segment size (segments with m tokens, influencing the granularity of the dispersion measure). We expect the segment size to be of particular importance, as choosing extreme values affects the calculations very strongly: using a segment size of 1 token is equivalent to relative term frequencies; using unsegmented texts is equivalent to document frequencies. Because Burrows (2007) gives no theoretical justification for his particular formulation of Zeta, a systematic exploration and evaluation is called for.

3. Text collection, code and raw data

Experiments have been performed using two very different text collections:

- A collection of French Classical and Enlightenment Drama (1630-1788): 150 comedies and 189 tragedies (from the Théâtre classique collection; Fièvre 2007-2017).
- A collection of Spanish novels (1880-1940): 24 novels from Spain and 24 from Latin America (from the CLiGS textbox: Henny 2017 and Calvo Tello 2017).

For reasons of space, we only report results for the Spanish novels. Texts, metadata, code, results and figures are available on Github: https://github.com/cligs/projects2018/tree/master/zeta-dh .

4. Methods and hypotheses

To obtain a better understanding of Zeta and its variants, we first visually explore the relation between segment size and the resulting zeta scores. We expect both Zeta variants and segment size to have visible consequences in this setting. Secondly, we evaluate the distinctiveness of words selected by different Zeta variants by using the highest ranked words as features in a classification task for distinguishing texts into two previously defined classes. This captures the degree to which the different Zeta variants and parameters identify words distinctive of these two classes. Note that we calculate Zeta scores from the complete set of documents. While this is not valid for a real-world classification task, it allows us to better judge the level of distinctiveness of the selected words. We expect better performance in the classification task with some of the new variants, compared to the classic “Burrows Zeta” (sd0). We also expect extreme segment lengths to significantly impact classification performance. We primarily aim at a methodological contribution here, so we do not attempt include a discussion of our results from a literary perspective (but see Schöch 2018 for such a contribution).

5. Exploratory approaches to Zeta variants and parameter variation

First, we take a closer look at the relationship between overall frequency and Zeta scores as it evolves with increasing segment size.

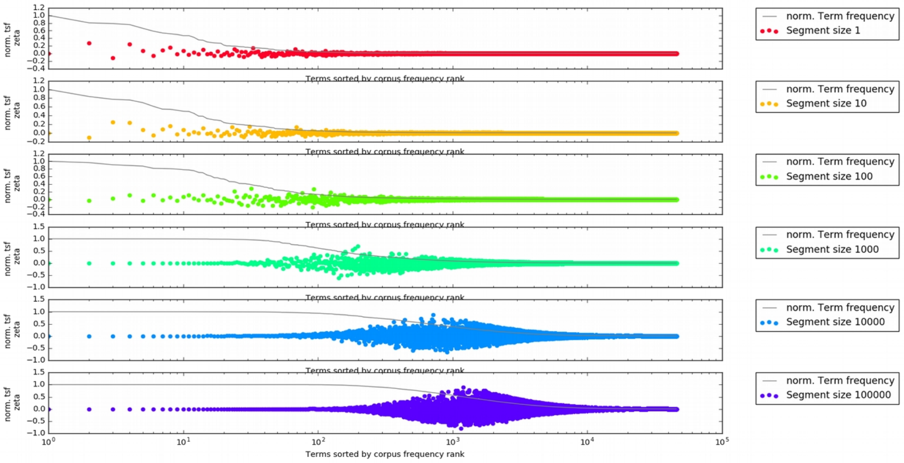

Figure 1. Distribution of Burrows Zeta (sd0) scores depending on segment size. Each dot is one word, ordered by descending frequency. The x-axis is log-scaled.

Figure 1 shows that when using very short segments, only highly frequent words (such as function words) can get high Zeta scores. With longer segments, words that are somewhat less frequent overall (well-interpretable content words) can also reach high Zeta scores; a desirable effect.

Additionally, we explore the influence of segment length and Zeta variant on the relation between segment proportions and zeta scores.

Figure 2 shows that with increasing segment size, Zeta scores generally increase because segment proportions increase. It also shows that in Burrows’ Zeta (left), terms with low segment proportions can never gain high Zeta scores. This limitation motivates the log2 and division variants that alleviate this effect: here, words to the bottom and left of the plots can also obtain extreme Zeta scores.

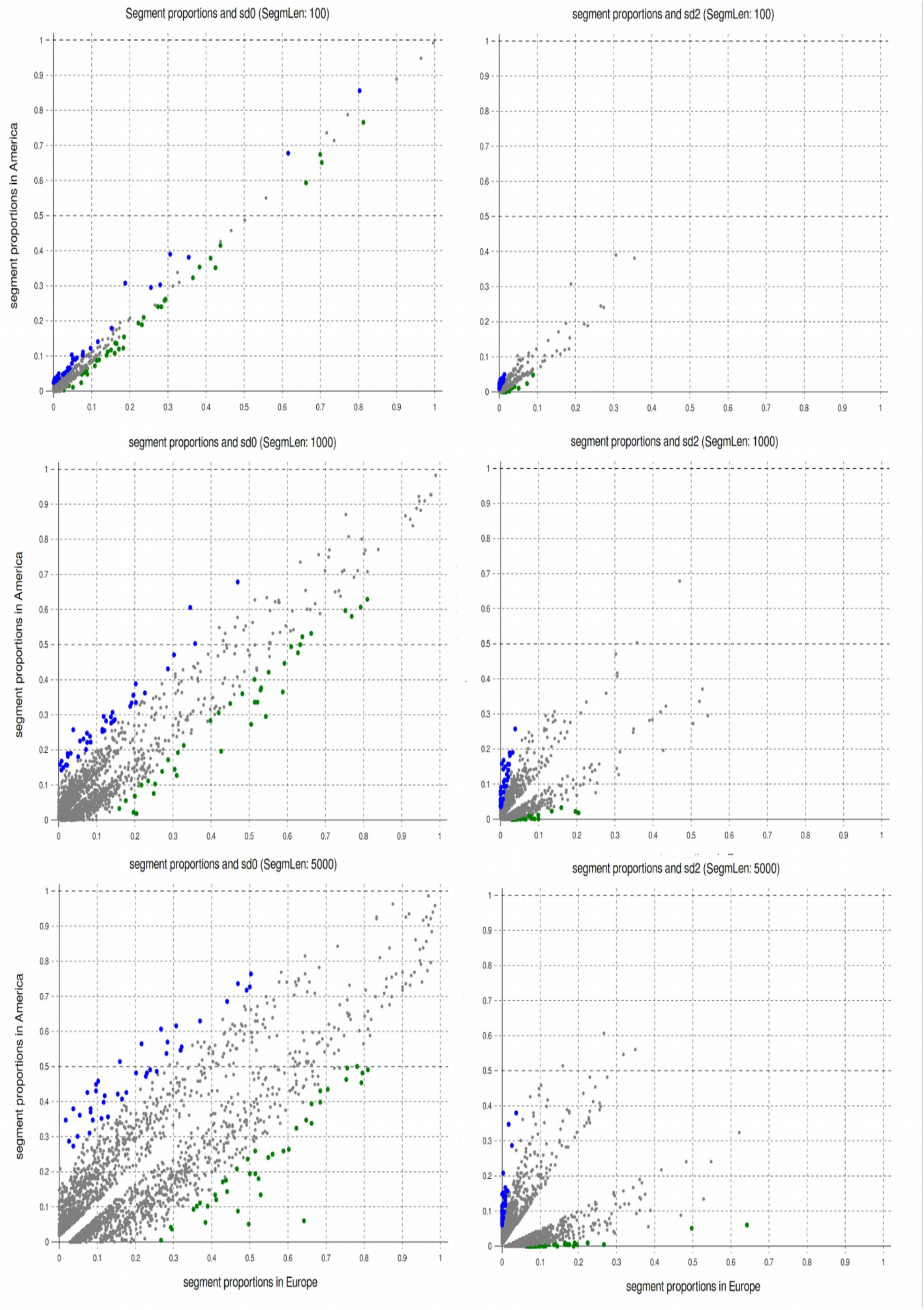

Figure 2. Segment proportions and Zeta scores for two Zeta variants (left: Burrows Zeta; right: log2-Zeta) and three segment sizes (100, 1000, 5000). Each dot is one word, 500 top Zeta words are shown. Colors indicate the words with the 40 highest (green) and lowest (blue) Zeta scores.

6. Evaluation of Zeta variants and parameters

Evaluating the performance of Zeta variants and different parameters is non-trivial, because it is impossible to define a human-annotated gold standard for distinctive words. Therefore, we use a classification task (with a Linear-SVM classifier and 3-fold cross-validation) for evaluation (the Spanish novels have to be classified by their continent of origin: America and Europe). The baseline of classification performance is F1=0.49 on average across all conditions and has been obtained using the top-80 most frequent words weighted with TF-IDF (see Robertson 2004).

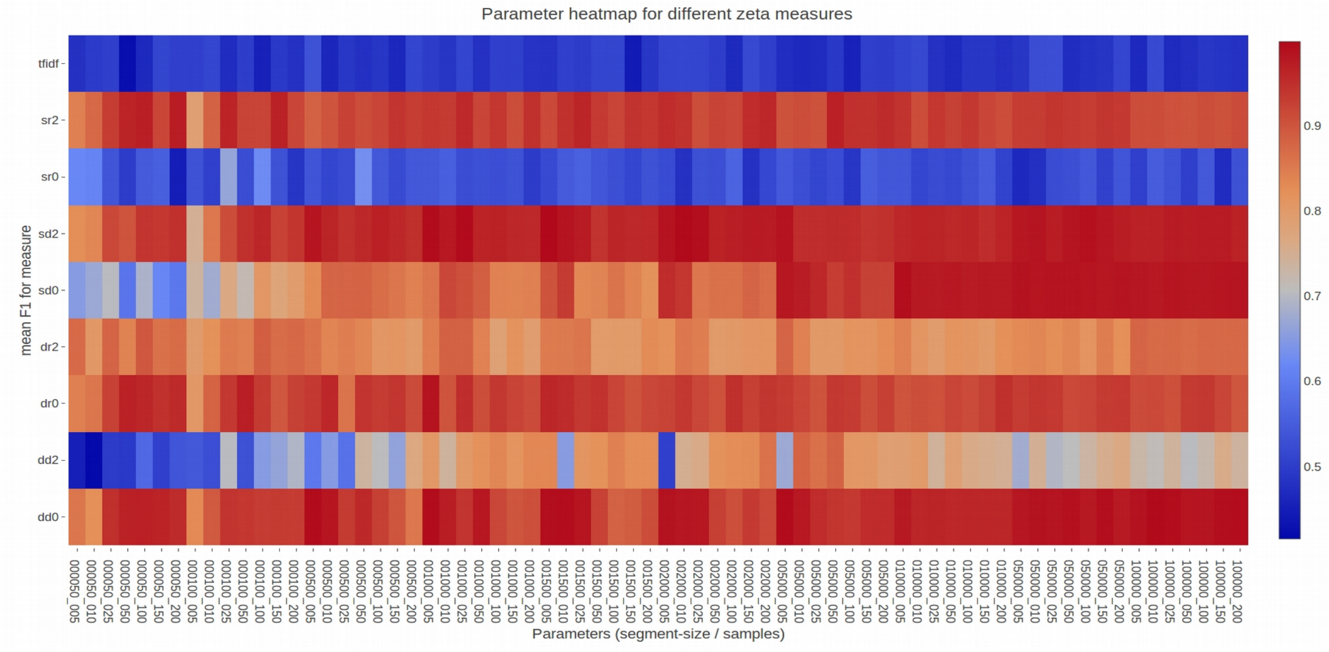

Figure 3. Classification performance depending on Zeta variant, segment size and number of segment-samples. We report mean F1-score over 15x3 folds.

Figure 3 shows that, as expected, most Zeta variants outperform the baseline. Segment size also influences performance: Burrows Zeta (“sd0”) performs particularly poorly with small segments (50, 100) and particularly well with large segments (>10000). Contrary to our expectation, large segment sizes do not generally have a negative impact on performance. The log2-Zeta variant (“sd2”) performs better than Burrows’ Zeta and is more robust with respect to segment size. In addition, we evaluate the parameter sampling size (number of segments randomly sampled for each document). For Burrows’ Zeta (“sd0”), we observe a better classification performance for small samples.

7. Discussion: Interpretability

While improved performance and robustness are welcome, another important characteristic of Burrows’ Zeta should not be forgotten, namely the high interpretability of the most distinctive words it identifies. The question is whether the gain in performance obtained with log2-Zeta comes at the expense of interpretability of the most distinctive words. Currently, we can merely offer some preliminary observations on this issue: First, the interpretability of distinctive words could be operationalized in a first approximation as the proportion of content words (nouns, verbs and adjectives) in the list of the most distinctive words, as opposed to function words and named entities.

Second, segment size and Zeta variant both appear to influence the types of words Zeta that determines to be distinctive: for example, very small segment sizes favor highly frequent function words, while very large segment sizes lead to place and person names taking up a considerable space in the word list. Also, some Zeta variants, including log2-Zeta, produce lists of words containing high proportions of place and person names even at intermediate segment lengths, that is wordlists that are less interpretable (see annex). These preliminary observations point to a possible trade-off between performance and interpretability that requires further, systematic investigation.

8. Conclusions and Future Work

Our experiments have allowed us to gain a much more detailed understanding of how Zeta works, mathematically and empirically. Additionally, we have identified at least one Zeta variant (“log2-Zeta”) that selects more distinctive words with regard to our classification task and is more robust against variation in segment length than Burrows Zeta.

As future work, we plan to conduct an investigation into the notion of “interpretability” and its relation to classification performance. Also, we plan to build an interactive visualization for our results to support a dynamic exploration of Zeta variants, key parameters and their influence on classification accuracy and distinctive words obtained. A larger agenda item is the evaluation of a substantial number of measures of distinctiveness, including Zeta, in a common framework.

9. Annex

For reasons of space, the wordlist annex can be found at:
https://github.com/cligs/projects2018/blob/master/zeta-dh/annex.pdf .

Bibliography

1. **Anthony, L.** (2005). AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit. *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning*. 7–13.
2. **Burrows, J.** (2007). All the Way Through: Testing for Authorship in Different Frequency Strata. *Literary and Linguistic Computing*, 22(1): 27–47 doi:10.1093/llc/fqi067.
3. **Calvo Tello, J.** (ed.) (2017). Corpus of Spanish Novel from 1880-1940. (CLiGS Textbox). Würzburg: CLiGS. https://github.com/cligs/textbox/tree/master/spanish/novela-espanola.
4. **Craig, H. and Kinney, A. F.** (eds). (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.
5. **Eder, M., Kestemont, M. and Rybicki, J.** (2016). Stylometry with R: A package for computational text analysis. *The R Journal*, 16(1): 1–15. https://journal.r-project.org/archive/2016/RJ-2016-007/index.html.
6. **Fièvre, P.** (ed). (2007). *Théâtre classique*. Paris: Université Paris-IV Sorbonne. http://www.theatre-classique.fr.
7. **Gries, S. T.** (2008). Dispersions and adjusted frequencies in corpora. I *nternational Journal of Corpus Linguistics*, 13(4): 403–37. doi:10.1075/ijcl.13.4.02gri.

8. **Heiden, S.** (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In Otoguro, R., et al. (eds), *24th Pacific Asia Conference on Language, Information and Computation - PACLIC24*. Sendai: Waseda University, 389–98. https://halshs.archives-ouvertes.fr/halshs-00549764/en.
9. **Henny, U.** (ed.) (2017). Collection of 19th Century Spanish-American Novels (1880-1916). (CLiGS Textbox). Würzburg: CLiGS. https://github.com/cligs/textbox/tree/master/spanish/novela-hispanoamericana.
10. **Hoover, D. L.** (2010). Teasing out Authorship and Style with t-tests and Zeta. *Digital Humanities Conference*. London http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-658.html.
11. **Kilgarriff, A.** (2001). Comparing Corpora. *International Journal of Corpus Linguistics*, 6(1): 97–133. doi:10.1075/ijcl.6.1.05kil.
12. **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.
13. **Lyne, A. A.** (1985). *Dispersion. The Vocabulary of French Business Correspondence*. Paris: Slatkine / Champion, pp. 101–24.
14. **Rayson, P. and Garside, R.** (2000). Comparing corpora using frequency profiling. *Proceedings of the Workshop on Comparing Corpora*. Hong Kong: ACM, 1–6.
15. **Robertson, S.** (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, 60(5): 503–20.
16. **Ruxton, G. D.** (2006). The unequal variance t-test is an underused alternative to Student’s t-test and the Mann–Whitney U test. *Behavioral Ecology*, 17(4): 688–90. doi:10.1093/beheco/ark016.
17. **Schöch, C.** (2018). Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie. In Bernhart, T., et al. (eds.), *Quantitative Ansätze in der Literatur- und Geisteswissenschaften*. Berlin: de Gruyter. 77-94. https://www.degruyter.com/viewbooktoc/product/479792.
18. **Schöch, C., Calvo Tello, J., Henny-Krahmer, U. and Popp, S.** (under review). The CLiGS textbox: Building and Using Collections of Literary Texts in Romance Languages Encoded in XML-TEI. *Journal of the Text Encoding Initiative*. Preprint: http://cligs.hypotheses.org/files/2017/09/Schoech-et-al_2017_Textbox.pdf.
19. **Scott, M.** (1997). PC Analysis of Key Words and Key Key Words. System, 25(2): 233–45.