Zeta & Eta: An Exploration and Evaluation of Two Dispersion-based Measures of Distinctiveness

Keli Du, Julia Dudar, Cora Rok and Christof Schöch

University of Trier, Germany

Abstract

In Corpus Linguistics, numerous statistical measures have been adopted to analyze large amounts of textual data in a contrastive perspective, in order to extract characteristic or "distinctive" features. While the most widely-used keyness measures are based on word frequency, an increasing number of research papers recently suggested dispersion-based measures as a better solution. These, however, are not new to Computational Literary Studies (CLS). In 2007, John Burrows introduced Zeta, a statistical measure that is mainly based on the degree of dispersion of a feature in a text corpus. In this paper, we also introduce *Eta*, a new measure of distinctiveness that is based on *deviation of proportions* suggested by Stefan Gries. By comparing Eta with Zeta, we demonstrate that both measures are able to identify relevant, interpretable distinctive words in a target corpus. Additionally, we make a first attempt to detect the key differences between these two measures by interpreting the top distinctive words.

Keywords

Computational Literary Studies, measure of distinctiveness, Zeta, Eta, dispersion

1. Introduction

In Linguistics and Literary Studies, comparing groups of texts – e.g. belonging to different literary genres or written for different audiences – is a fundamental procedure [11, see e.g.,]. In Corpus Linguistics, numerous statistical measures and instruments have been introduced and adopted for investigating and analyzing large amounts of textual data in a contrastive perspective [e.g. 20, 17, 15]. They are usually referred to as 'keyness measures', as they operate on a lexical level and are used for extracting "key" terms or phrases. We prefer the term 'measures of distinctiveness', as it better emphasizes that this kind of analysis is about the extraction of characteristic words on the basis of a comparison [see 24].

The most widespread keyness measures used in Corpus Linguistics are frequency-based – for example, the chi-squared test or the log-likelihood-ratio test [25], implemented e.g. in AntConc [1]. Recently, several research papers suggested dispersion-based measures as a better solution for contrastive corpus analysis [e.g. 4, 8, 7]. Apart from that, the use of dispersion in the search for important text features is not new to Computational Literary Studies (CLS). In 2007, John Burrows introduced Zeta, a keyness measure that is mainly based on the degree of dispersion of a feature in a text corpus [2]. Originally, it was used in the context of authorship

CHR 2021: Computational Humanities Research Conference, November 17–19, 2021, Amsterdam, The Netherlands

duk@uni-trier.de (K. Du); dudar@uni-trier.de (J. Dudar); rok@uni-trier.de (C. Rok); schoech@uni-trier.de (C. Schöch)

0000-0001-7800-0682 (K. Du); 0000-0001-5545-9562 (J. Dudar); 0000-0001-9698-7513 (C. Rok); 0000-0002-4557-2753 (C. Schöch)

attribution, but it later came to be used also to solve other issues in CLS, including corpus comparison [e.g. 3, 9, 23].

There are several important studies that explore and evaluate frequency-based measures [e.g. 10, 18, 12, 19, 6], and some studies that compare dispersion based measures to frequency based measures [e.g 4, 8, 12]. However, as far as we know, no attempt has been made to compare the dispersion-based measures to each other. In our project "Zeta and company"[1] we aim to enhance the understanding of both frequency- and dispersion-based measures by implementing them in a Python framework. Based on tests with literary texts we evaluate which measures perform best for different tasks and kinds of textual data. This article presents a pilot study in our project and it aims to perform a statistical analysis and a qualitative evaluation of two dispersion-based distinctiveness measures: (1) Eta, which is based on deviation of proportions (DP), developed by Stefan Gries; (2) Zeta, which was proposed by John Burrows.[2]

Firstly, we will explain how Eta and Zeta are calculated. After that, using a collection of 160 novels of four different subgenres published in France in the 1980s, we will examine how Eta behaves in contrast to Zeta and how their relationship changes when the segment length varies. The following questions will be addressed: How useful is Eta as a basis for identifying distinctive words in one text group compared to another text group? What are the differences between Eta and Zeta and what results do they display?

2. Keyness analysis: from frequency to dispersion

Despite the dominance of frequency-based keyness measures (e.g. chi-squared test, log-likelihood ratio test), there are several alternative measures which consider other types of information like the distribution of words (e.g. t-Test, Mann-Whitney-U-test) and their dispersion (e.g. Zeta). A helpful overview of the frequency- and distribution-based measures can be found in [12]. In addition, Machine Learning-approaches (e.g. weights of a linear SVM) or entropy-related approaches (e.g. Kullback-Leibler divergence, see [5]) can be used to identify distinctive words in a target corpus.

As already mentioned, the most widely used keyness measures in Corpus Linguistics are frequency-based and they do not consider how the particular words are distributed within a corpus. This means that a word can be marked as distinctive for the entire target corpus, even if it just appears very frequently in a small number of texts. For illustration, Figure 1 presents the result of an analysis carried out using AntConc's log-likelihood ratio test on our working corpus (described below): keywords where extracted from a comparison of 40 French science fiction novels (as the target corpus) with 120 French novels of other subgenres (as the comparison corpus).[3] It turns out that the top-ranked words are almost entirely proper names. Each of them appears only in one novel of the target corpus, albeit very frequently, and likely not at all in the comparison corpus and therefore cannot truly represent the entire target corpus. In order to obtain more meaningful results, proper names should be pruned from the list.

To deal with this challenge, the dispersion of a feature, which is the degree of an even distribution of a feature, should be considered as well (on dispersion, see [13]; for the use

[1] See: https://zeta-project.eu/en/.

[2] We have implemented both measures in our Python framework. See: https://github.com/Zeta-and-Company/pydistinto.

[3] AntConc 3.5.9 [see 1] was used with the following keyness parameters: Log-Likelihood (4-way) and a p-value cut-off of 0.001. The measure of effect size shown is DIFF.

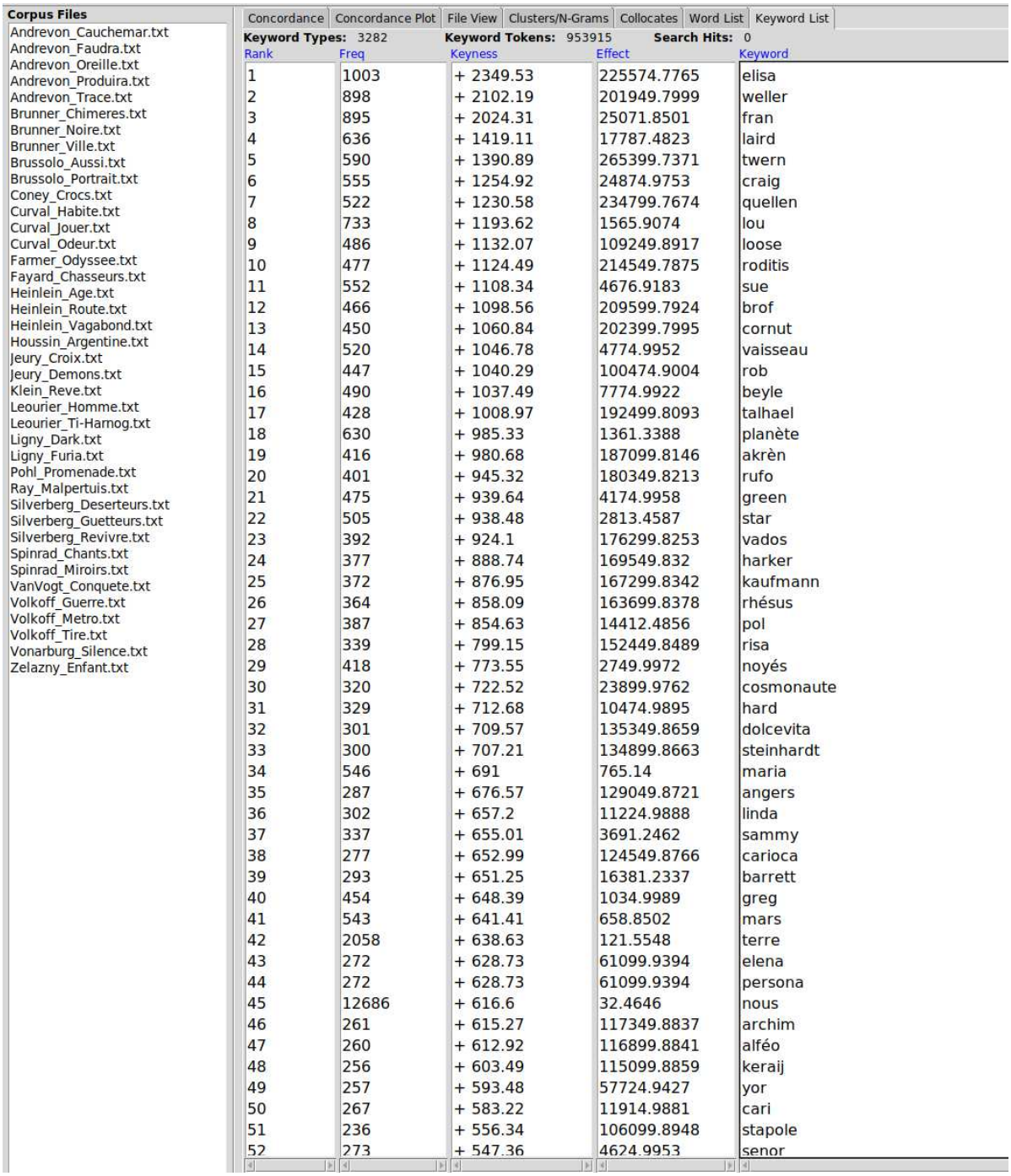


Figure 1: Log-likelihood ratio test with AntConc.

of dispersion for keyness analysis, see [4]). Gries [8] gives a detailed overview of dispersion measures and proposes his own measure, called deviation of proportions (DP).

DP compares the difference between observed and expected relative frequency of a word in every single document of the corpus in order to quantify the dispersion of the word:

DP is calculated as follows: for each corpus part (e.g., a file), compute $\boldsymbol{s}$, which represents how much of the corpus it constitutes (as a fraction of the whole corpus) and $\boldsymbol{v}$, which represents how much of the word in question it contains (as a fraction of the word's frequency). Then subtract all $\boldsymbol{s}$-values from all $\boldsymbol{v}$-values, take the absolute values of those differences, sum them up, and divide by two [7].

$$DP = \frac{\sum_{i=1}^{n} |s_i - v_i|}{2}$$

The theoretical range of DP values is between 0 and 1. A value of 0 reflects a perfectly even dispersion, while a value of 1 represents a maximally uneven dispersion. This measure seems to have several advantages compared to other dispersion measures. For example, it can handle corpus parts of different lengths and it can distinguish between slight variations in distribution without being overly sensitive. However, there is still a lack of empirical evidence supporting the use of DP.

As mentioned before, Burrows' Zeta also considers dispersion and it is calculated by comparing the document proportion (docP) of each feature in the target and in the comparison corpus. At first, each text in each group is divided into segments of a certain length (segment length is a key parameter of the measure). For each word $\boldsymbol{w}$ in the vocabulary, docP is calculated by establishing the proportion of segments in which the word occurs at least once, so docP ranges between 0 and 1.

In order to find out whether a word is distinctive for the target coups, the docP or devP[4] values of the word in the target and the comparison corpus must be compared, respectively. Based on docP and devP, two measures of distinctiveness can be defined. The Zeta score of (w) is the subtraction of docP in the comparison corpus from that in the target corpus [see 21]. Therefore, the theoretical range of the Zeta score is between -1 and 1. The words with the highest Zeta scores are the most distinctive words of the target corpus. By analogy, and using devP instead of docP as the measure of dispersion, a new measure of distinctiveness can be defined, which we call Eta. It is obtained by subtracting the devP of a word (w) in the comparison corpus from the devP of the same word in the target corpus. Contrary to docP, a small devP of a word reflects a more even distribution of a feature in a corpus. It is therefore expected that the devP of distinctive words in the target corpus is smaller than the devP of these words in the comparison corpus. So the words with the lowest Eta scores are the most distinctive words of the target corpus.[5] As we can see here, although Zeta and Eta are both dispersion-based measures, they have a different mathematical definition of dispersion. As Eta takes into account the ratio of document size and corpus size, which Zeta doesn't, we intend to test whether or not Eta performs better in detecting distinctive words than Zeta.

3. Tests and results

3.1. Corpus

The corpus used in this study is a collection of 160 novels published in France between 1980 and 1989. 120 of them are lowbrow novels of three subgenres (40 novels for each subgenre): sentimental novels, crime fiction and science fiction. The remaining 40 are highbrow novels.

[4] We use devP instead of DP to better distinguish between the two terms.

[5] Only words which appear at least once in both corpora will be considered here and in the following, because devP does not yield meaningful results otherwise.

The corpus size is approximately nine million words. All texts have been lemmatized using Treetagger and the units of calculation are lemmas. As our goal was to extract distinctive lemmas for each subgenre, we used a one-vs-rest strategy: the target corpus contains 40 novels of one subgenre and the comparison corpus contains 120 novels of the other three subgenres. This allowed us to focus on extracting distinctive features that are strongly related to the unique characteristics of the target corpus.[6]

3.2. Statistical observations

The results of our comparative analysis are two lists of words which are ranked by their Zeta or Eta scores, respectively. To compare the differences of Zeta and Eta, we measure the ranking correlation between the two word lists using Spearman's rank correlation. The stronger the correlation, the less different these two word lists are. We performed tests on four comparison groups: sci-fi vs. non-sci-fi, etc. for each genre. The results of these four tests were almost the same. For illustration, the results presented below are based on the comparison of sci-fi vs. non-sci-fi.

As it is common to split novels into segments when applying Zeta, we also wanted to examine the impact of the segment size on the results. So we did our tests using three segmentation strategies: split all novels into (1) 5000-word segments, (2) 10000-word segments and (3) take each novel as a segment without chunking. (The median length of the novels is about 46800 words.) For (1) and (2), segments shorter than 5000 or 10000 were removed from the corpus.

Before comparing Zeta and Eta, we first compared the underlying values: the docP and the devP. Again, Spearman's correlation between the word rankings based on these two dispersion measures was analyzed. In both corpora, the ranking correlations of the three tests with different segment length are -1, -1, and -0.98, respectively. Figure 2 illustrates the relation between docP and devP for all words in the target corpus.[7] Each blue point represents a word and the three graphs from left to right show the results of the tests on 5000-word segments, 10000-word segments and novel segments without chunking, respectively. Clearly, devP and docP have a strong negative correlation, but the distribution of points in the three graphs from left to right becomes increasingly dispersed. This means that the longer the novel segments are, the less similar the word list rankings between devP and docP are.

The comparison of Zeta and Eta leads to identical results. The strong negative correlations between the word rankings in the three tests are -0.99, -0.99, and -0.85, respectively. Each blue point in Figure 3 represents a word and the x and y axes are the Zeta and Eta scores for each word. The three graphs from left to right show the results of tests on 5000-word segments, 10000-word segments and entire novels, respectively. We can observe that the distribution of points gradually becomes more dispersed. This means that the longer the novel segments are, the less similar the Zeta and Eta scores are.

Comparing the top distinctive words found by Zeta and Eta for each subgenre, we can often observe the same words, but in a different order. To quantify these differences, we calculated the token based Jaccard similarity and NLTK's edit distance between the top ten to 500 Zeta

[6]The texts contained in the corpus are in-copyright texts that we are using in the framework of the "Text and Data Mining Exception" defined in German copyright law (§60d Urhg), following the EU "Directive on Copyright in the Digital Single Market". While the corpus cannot be shared as it is, we plan to publish derived features [see 22] that allow others to repeat our calculations.

[7]The scatter plot of docP and devP of words in the comparison corpus is almost the same as that in the target corpus, so it will not be displayed again.

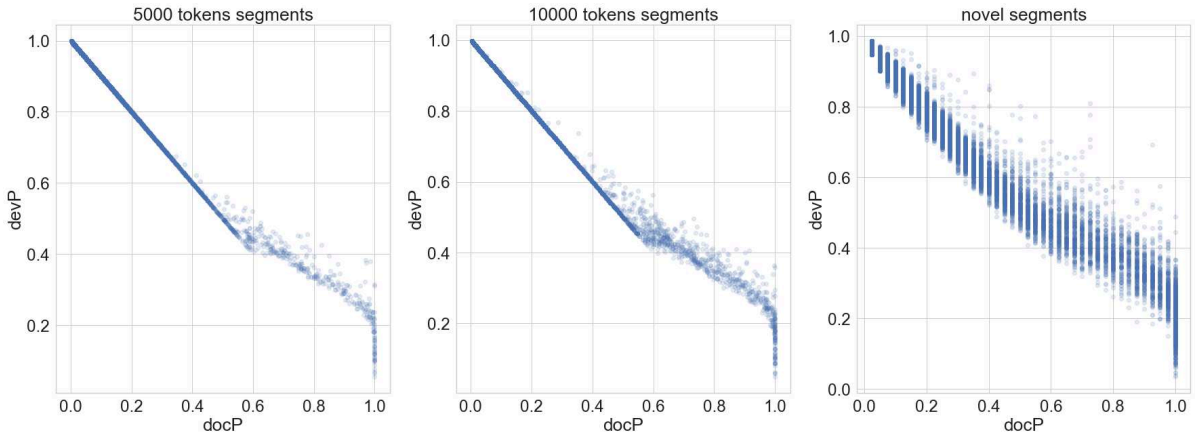


Figure 2: Scatter plot of docP and devP of words in the target corpus.

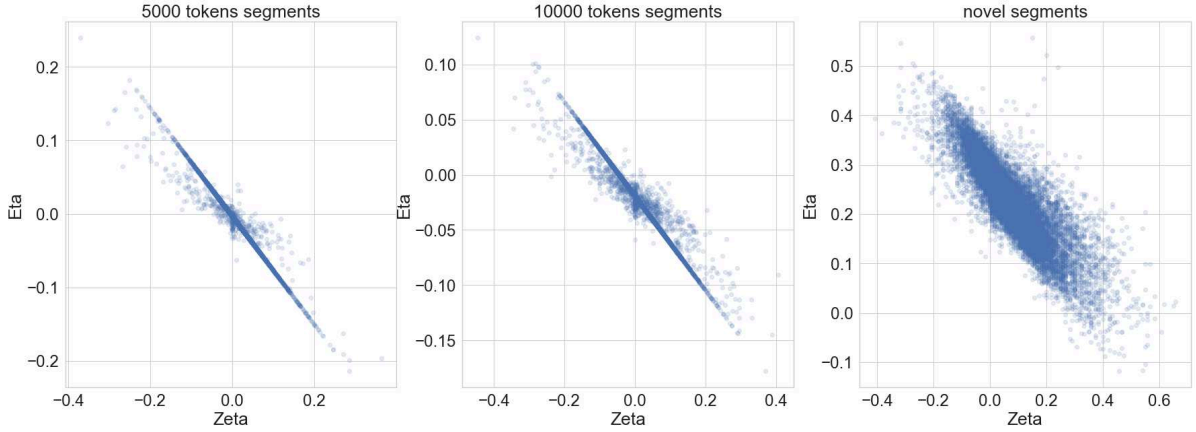


Figure 3: Scatter plot of Zeta and Eta.

and Eta words for different segment lengths.[8] In Figure 4, the first and the second row are the Jaccard similarity results and the NLTK's edit distance results, respectively. The four columns are the results of each of the four subgenres (from left to right: highbrow, crime, sci-fi and sentimental) taken as a target corpus. The results of both Jaccard similarity and NLTK's edit distance show an increasing trend. The increase of the Jaccard similarity indicates that, as the number of top words increases, the overlap of the Zeta and Eta word lists increases gradually. Splitting novels into shorter segments leads to a greater overlap. In contrast to this result, the increase of the NLTK's edit distance shows that the words are ranked more differently with the increase of the number of top words. These observations also prove our earlier point: the shorter the segments, the more words have the same or similar rank in both lists.

[8]The Jaccard similarity [see 16] calculates the size of the intersection divided by the size of the union of two word lists without considering the ranking of words. Larger values indicate a greater overlap between the top Zeta and Eta words. In contrast to the Jaccard similarity, the NLTK's edit distance (https://www.nltk.org/api/nltk.metrics.html#nltk.metrics.distance.edit_distance, see Levenshtein edit-distance, [14]) takes the ranking of words into consideration and counts the number of words that need to be substituted, inserted, or deleted, to transform one list into another. Larger values indicate a greater difference between the Zeta and Eta word lists.

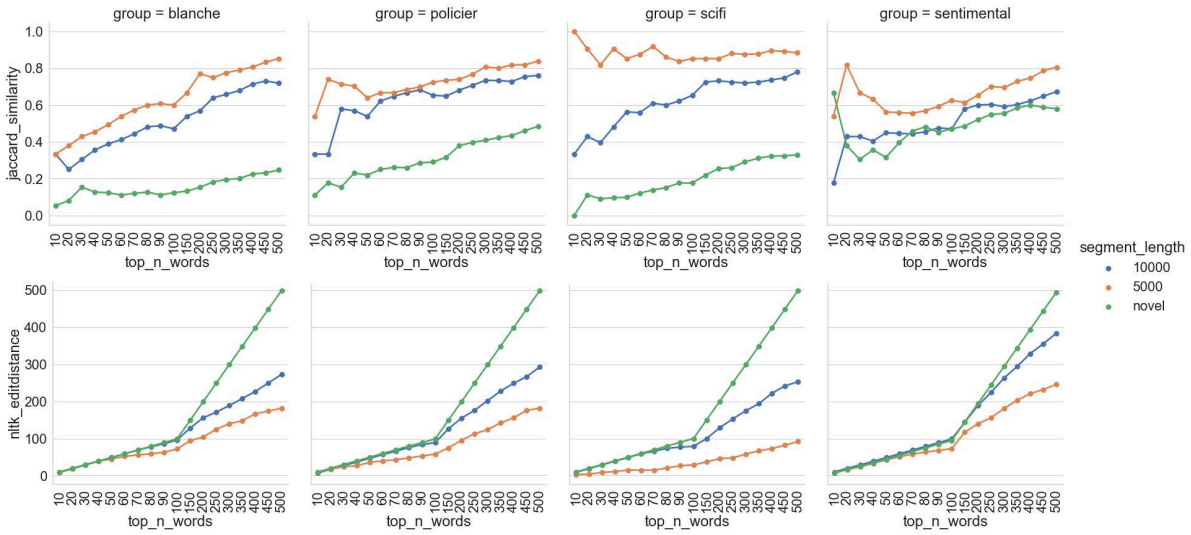


Figure 4: Jaccard similarity (top row) and NLTK's edit distance (bottom row) between the top 10 to 500 Zeta- and Eta-words, for three segment lengths.

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

Figure 5: Top ten Zeta (left) and Eta (right) words of a 5000-word segment analysis.

3.3. Interpretation of the word lists

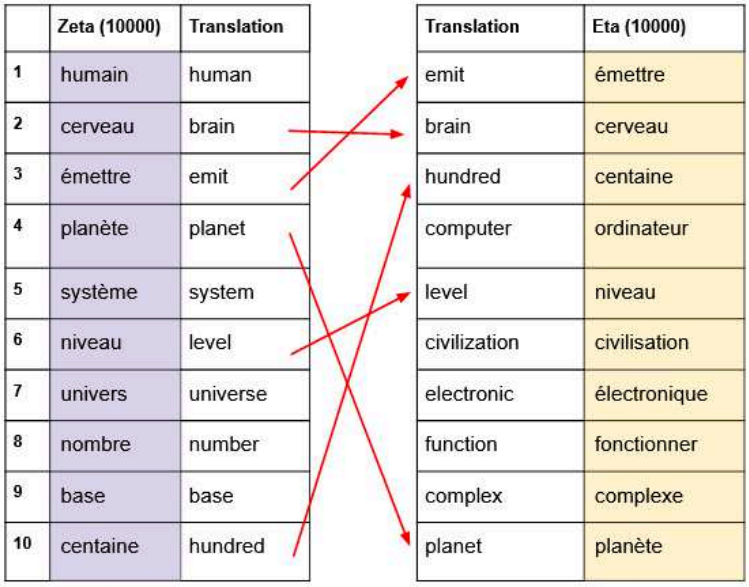
Figure 5 shows the top ten distinctive Zeta and Eta words of the science fiction corpus split into 5000-word segments. Both word lists contain the same genre-specific words with a slightly different ranking.

To better illustrate the results of the different tests, we assigned the words to semantic categories. Figure 6 shows the (heuristic) categorization of the words of the first test.

Figure 7 shows the results of the analysis with 10000-word segments: there are only five

	SEMANTIC CATEGORIES	ZETA | ETA
1	lifeform	human, brain
2	space	planet, universe
3	spatial data	center
4	computation	number
5	technology	system, screen
6	movements	achieve, emit

Figure 6: A heuristic categorization of the top ten words of the 5000-word segments analysis.



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

Figure 7: Top Ten Zeta and Eta words of a 10000-words segment analysis.

overlapping words in the top 10 words. The top 30 Zeta words, however, contain more of the highly ranked Eta words than vice versa.

If we compare the two Zeta word lists in Figures 5 and 7, we notice that the Zeta words do not change much with the increased segment length: There are three new words in the top ten list, “level”, “base” and “hundred”, whereas the words “human”, “brain”, “planet”, “universe”, “number”, “system” and “emit” can already be found in the first Zeta word list, which indicates a certain consistency. The Eta word list in turn displays more new distinctive words (“civilisation”, “level”, “complex”, “hundred”, “computer”, “function”, “electronic”). However, the words of both lists can be assigned to the previously defined semantic categories (Figure 8).

Figure 9 shows the word lists of our third analysis, where a whole novel represents a segment.

	SEMANTIC CATEGORIES	ZETA	ETA
1	lifeform	human, brain	brain, civilization
2	space	planet, universe	planet
3	spatial data	level, base	level, complex
4	computation	number, hundred	hundred
5	technology	system	computer, function, electronic
6	movements	emit	emit

Figure 8: A heuristic categorization of the top ten words of the 10000-word segments analysis (the words in yellow are new compared to the 5000-word segment analysis).

	Zeta (novel)	Translation
1	orbite	orbit
2	civilisation	civilization
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	Eta (novel)
partial	partiel
chemical	chimique
functioning	fonctionnement
broadcasting	diffusion
diameter	diamètre
hypnotic	hypnotique
radiation	radiation
criterion	critère
govern	régir
vertebral	vertébral

Figure 9: Top ten Zeta and Eta words of the novel as a segment analysis.

It is noticeable that there is no intersection between the words of both lists; only two of the top ten words of each list can be found in the other, namely under the top 25 (Eta rank 14: "concept"; Eta rank 23: "nuclear" / Zeta rank 19: "chemical"; Zeta rank 14: "functioning").

While the Zeta list contains words like "humanity", "civilization", "space", "orbit", "earthly", "computer", "electronic" and "robot", which seem to fit into the previously established semantic categories and represent more general terms from everyday language, the Eta words like "diameter" or "vertebral" are more specific and sophisticated and open up further semantic categories from the fields of science (Figure 10). This tendency of extracting more new specific words by Eta becomes even stronger when the segment length increases up to novel length, while the Zeta words stay more general. As Eta words seem more specific, our assumption is that they should be less frequent than the Zeta words in a much larger corpus. To verify this, we checked the frequency of the top Zeta and Eta words in the French Wikipedia.[9] Figure 11

[9]The frequency of words in Wikipedia are obtained from http://redac.univ-tlse2.fr/corpora/wikipedia_en.

	SEMANTIC CATEGORIES	ZETA	ETA
1	lifeform	humanity, civilization	govern
2	space	orbit, earthly	
3	computation		partial, diameter
4	technology	computer, electronic, robot	functioning
5	physics	magnetic, nuclear	radiation
6	chemistry		chemical, diffusion
7	anatomy		vertebral
8	psychology		hypnotic
9	theories	concept	criterion

Figure 10: A heuristic categorization of the top ten words of the novel as a segment analysis (the categories in yellow are the 'new' ones, established for the third analysis).

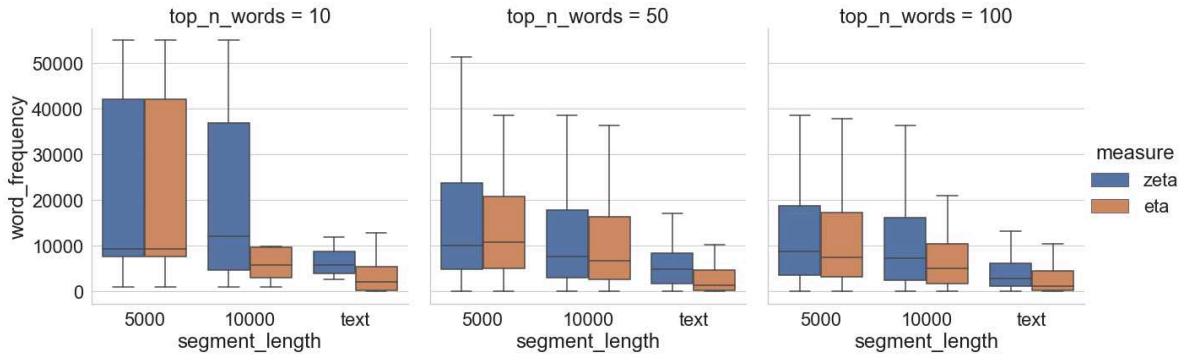


Figure 11: Word frequency of top Zeta and Eta words in French Wikipedia.

shows that the top (10, 50 and 100) Zeta words are indeed more frequent and therefore less specific than the Eta words. This effect is stronger, the longer the segment length is.

html. If a word doesn't exist in the frequency table, the frequency is set to 0.

4. Conclusion and future work

This paper presents a comparison of two measures of distinctiveness, Zeta and Eta. The results show that on the statistical level, both of them have a very strong negative correlation, despite their different basis for calculation. Another observation is that the correlation between Zeta and Eta is stronger when novels are divided into shorter segments. We obtain the weakest correlation when novels are not split into segments at all. This correlation is also reflected in the word lists: the shorter the segments, the more similar the word lists and vice versa. The calculation of the Jaccard similarity allowed us to observe the following trend: The Jaccard similarity decreases, when the segment length increases.

The observed similarities concern word rankings as well: We observe not only (almost) the same words in the top ten ranking when calculating with small segments, but the word-rankings are also almost the same in both word lists. The calculation of the NLTK's edit distance between word lists verified our observation: The distance between the word-rankings increases when the segment length increases.

A qualitative interpretation of the word lists confirmed the statistical observations. Both measures are able to identify relevant interpretable distinctive words in a target corpus. There is no need to use stop words or to prune proper names: Both dispersion-based measures mark content words as distinctive. It seems that when the segment length increases, the Zeta words remain content-related and more general, while the Eta words also remain content-related, but become more specific. We are going to investigate this phenomenon in further tests.

In the future, we plan to deepen our understanding of distinctiveness measures even further. Our next steps are to test the measures on larger and more varied corpora and make more experiments with segment length. We are also planning to include other distinctiveness measures in our framework, such as Kullback-Leibler Divergence, Wilcoxon signed-rank test or T-test. One point to emphasize is that the qualitative interpretation of the word lists may seem very subjective and it looks more like an exploration than an evaluation. This is inevitable, because as far as we know, a widely accepted robust method for a qualitative evaluation in this area is still lacking. Therefore, we will work on developing new evaluation strategies for these measures, in order to explore the advantages and disadvantages of each of these measures and to find out for which purpose they should be used.

Author contributions

All authors contributed to the conceptualization of the research, investigation, formal analysis, writing the original draft and editing and reviewing the text. Specific additional contributions: KD contributed to project administration, software development, visualisation and methodology. JD contributed to data curation and software development. CR contributed validation. CS contributed to data curation, software development, funding acquisition and supervision. Author order is alphabetical. All authors gave final approval for publication and agree to be held accountable for the work performed therein.[10]

[10] See https://casrai.org/credit.

References

[1] L. Anthony. "AntConc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom". In: 2005, pp. 729–737. DOI: 10.1109/ipcc.2005.1494244.

[2] J. Burrows. "All the Way Through: Testing for Authorship in Different Frequency Strata". In: *Literary and Linguistic Computing* 22.1 (2007), pp. 27–47. DOI: 10.1093/llc/fqi067. URL: http://llc.oxfordjournals.org/content/22/1/27.abstract.

[3] H. Craig and A. F. Kinney, eds. *Shakespeare, Computers, and the Mystery of Authorship*. 1st ed. Cambridge University Press, 2009.

[4] J. Egbert and D. Biber. "Incorporating text dispersion into keyword analyses". In: *Corpora* 14.1 (2019), pp. 77–104. DOI: 10.3366/cor.2019.0162. URL: https://www.euppublishing.com/doi/abs/10.3366/cor.2019.0162.

[5] P. Fankhauser, J. Knappen, and E. Teich. "Exploring and Visualizing Variation in Language Resources". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Ed. by N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014.

[6] C. Gabrielatos. "Keyness Analysis: nature, metrics and techniques". In: *Corpus Approaches to Discourse: A Critical Review* (2018), pp. 225–258. URL: https://research.edgehill.ac.uk/en/publications/keyness-analysis-nature-metrics-and-techniques-2.

[7] S. Gries. "A new approach to (key) keywords analysis: Using frequency, and now also dispersion". In: *Research in Corpus Linguistics* 9 (2021), pp. 1–33. DOI: 10.32714/ricl.09.02.02.

[8] S. T. Gries. "Dispersions and adjusted frequencies in corpora". In: 2008. DOI: 10.1075/ijcl.13.4.02gri.

[9] D. L. Hoover. "Teasing out Authorship and Style with t-tests and Zeta". In: *Digital Humanities Conference*. London, 2010. URL: http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-658.html.

[10] A. Kilgarriff. "Comparing word frequencies across corpora: Why chi-square doesn't work, and an improved LOB-Brown comparison". In: *ALLC-ACH Conference*. 1996, pp. 169–172.

[11] S. Klimek and R. Müller. "Vergleich als Methode? Zur Empirisierung eines philologischen Verfahrens im Zeitalter der Digital Humanities [Abstract]". In: *JLT Articles* 9.1 (2015). URL: http://www.jltonline.de/index.php/articles/article/view/758.

[12] J. Lijffijt, T. Nevalainen, T. Säily, P. Papapetrou, K. Puolamäki, and H. Mannila. "Significance testing of word frequencies in corpora". In: *Digital Scholarship in the Humanities* 31.2 (2014), pp. 374–397. DOI: 10.1093/llc/fqu064. URL: http://dsh.oxfordjournals.org/lookup/doi/10.1093/llc/fqu064.

[13] A. A. Lyne. "Dispersion". In: *The Vocabulary of French Business Correspondence: Word Frequencies, Collocations and Problems of Lexicometric Method*. Paris: Slatkine, 1985, pp. 101–124.

[14] G. Navarro. “A guided tour to approximate string matching”. In: *ACM Computing Surveys* 33.1 (2001), pp. 31–88. DOI: 10.1145/375360.375365. URL: https://dl.acm.org/doi/10.1145/375360.375365.

[15] M. L. Newman, C. J. Groom, L. D. Handelman, and J. W. Pennebaker. “Gender differences in language use: An analysis of 14,000 text samples”. In: *Discourse Processes* 45.3 (2008), pp. 211–236.

[16] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu. “Using of Jaccard coefficient for keywords similarity”. In: *Proceedings of the international multiconference of engineers and computer scientists*. Vol. 1. 2013, pp. 380–384.

[17] M. P. Oakes and M. Farrow. “Use of the Chi-Squared Test to Examine Vocabulary Differences in English Language Corpora Representing Seven Different Countries”. In: *Literary and Linguistic Computing* 22.1 (2007), pp. 85–99. DOI: 10.1093/llc/fql044. URL: https://academic.oup.com/dsh/article/22/1/85/1025876.

[18] M. Paquot and Y. Bestgen. “Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction”. In: *Corpora: Pragmatics and Discourse*. Ed. by A. H. Jucker, D. Schreier, and M. Hundt. Brill | Rodopi, 2009. DOI: 10.1163/9789042029101_014. URL: https://brill.com/view/book/edcoll/9789042029101/B9789042029101-s014.xml.

[19] P. Pojanapunya and R. W. Todd. “Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis”. In: *Corpus Linguistics and Linguistic Theory* 14.1 (2018), pp. 133–167. DOI: 10.1515/cllt-2015-0030. URL: https://www.degruyter.com/view/journals/cllt/14/1/article-p133.xml.

[20] P. Rayson, G. N. Leech, and M. Hodges. “Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus”. In: *International Journal of Corpus Linguistics* 2.1 (1997), pp. 133–152.

[21] C. Schöch. “Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie”. In: *Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven*. Ed. by T. Bernhart, S. Richter, M. Lepper, M. Willand, and A. Albrecht. Berlin: de Gruyter, 2018, pp. 77–94. URL: https://www.degruyter.com/view/books/9783110523300/9783110523300-004/9783110523300-004.xml.

[22] C. Schöch, F. Döhl, A. Rettinger, E. Gius, P. Trilcke, P. Leinen, F. Jannidis, M. Hinzmann, and J. Röpke. “Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen”. In: *Zeitschrift für digitale Geisteswissenschaften (ZfdG)* 5 (2020). DOI: http://dx.doi.org/10.17175/2020_006. URL: http://www.zfdg.de/2020%5C%5F006.

[23] C. Schöch, D. Schlör, A. Zehe, H. Gebhard, M. Becker, and A. Hotho. “Burrows’ Zeta: Exploring and Evaluating Variants and Parameters”. In: *Book of Abstracts of the Digital Humanities Conference*. Mexico City: Adho, 2018. URL: https://dh2018.adho.org/burrows-zeta-exploring-and-evaluating-variants-and-parameters/.

[24] J. Schröter, K. Du, J. Dudar, C. Rok, and C. Schöch. “From Keyness to Distinctiveness – Triangulation and Evaluation in Computational Literary Studies”. In: *Journal of Literary Theory (JLT)* ().

[25] M. Scott. “PC Analysis of Key Words and Key Key Words”. In: *System* 25.2 (1997), pp. 233–245.