jcls. Journal of Computational Literary Studies

OPEN ACCESS

Article

Exploring Measures of Distinctiveness
An Evaluation Using Synthetic Texts

Julia Havrylash[1]
Christof Schöch[1]

1. Trier Center for Digital Humanities, Trier University, Trier, Germany.

Citation
Julia Havrylash and Christof Schöch (2025). "Exploring Measures of Distinctiveness. An Evaluation Using Synthetic Texts". In: *Journal of Computational Literary Studies* 4 (1). 10.48694/jcls.4209

Date published 2025-11-21
Date accepted 2025-10-26
Date received 2025-02-06

Keywords
evaluation, measures of distinctiveness, keyness, synthetic texts

Reviewers
Erik Ketzan, Sebastian Padó

Note
This paper has passed through the conference track of JCLS. In addition to being peer reviewed, it was presented and discussed at the 4th Annual Conference of Computational Literary Studies at Krakow, Poland, in July 2025.

Abstract. Measures of distinctiveness (aka keyness) are important tools for comparing groups of texts to identify each group's characteristic features. Evaluating these measures is essential to ensure their reliability and predictability. In our research, we developed and applied a new method for evaluating measures of distinctiveness. Our method uses a synthetically generated, homogeneous text corpus to which we insert an artificial word whose frequency and dispersion are precisely manipulated. This approach allows us to determine each measure's sensitivity to variations in frequency and dispersion. Through our evaluation, we have uncovered previously unknown characteristics of these measures. Specifically, we discovered that the TF-IDF-based measure we used is more sensitive to dispersion variations than other dispersion-based measures. Moreover, we found that Eta cannot detect a word with a clear dispersion contrast when it has the same frequency in both the target and comparison groups. In our next steps, we aim to explore practical applications of this new knowledge about measures of distinctiveness.

1. Introduction

Comparing groups of texts to identify what is distinctive about each is a fundamental approach in many research contexts. In computational literary studies, such comparisons are particularly valuable for exploring literary style, genre conventions, authorial voice, or historical shifts in discourse. Such contrastive analyses of literary corpora have, for instance, been used to study different characters' speech in Shakespeare's plays (Culpeper 2009), in the context of gender markers in English-language novels (Weidman and O'Sullivan 2018), for determining the place of tragicomedy with respect to comedy and tragedy (Schöch 2018), or for identifying phrasal expressions characteristic of subgenres of the contemporary French novel (Gonon et al. 2018).

A key challenge in this task, alongside selecting the appropriate comparison corpora, is finding the most suitable measure and parameters for a specific research question and corpus composition. There is a wide range of measures available, and the list of most distinctive features they identify can vary considerably (as shown e.g. by Du et al. 2021a for the Zeta and Eta measures). While in principle, virtually any countable feature of texts may be submitted to a contrastive statistical analysis in order to identify distinctive features, we focus exclusively on lexical features in this research, specifically on word unigrams. In this paper, we explore and evaluate various measures of distinctiveness, also known as keyness measures, which support such research from a quantitative

perspective. Although we do not prescribe a particular measure for researchers to use, our paper offers valuable insights into the characteristics of these measures, helping researchers understand their behavior and the potential outcomes when applying different distinctiveness measures in their studies.

In the research we report on here, we focus on evaluating measures of distinctiveness through an analysis based on synthetic texts.[1] Our research proposes a new method for evaluating measures of distinctiveness, utilizing synthetically created text collections that reflect word frequencies as they would have occurred in a regular corpus built from the same original texts. Studies based on naturally occurring language must work around the fact that the frequency and dispersion of any word will vary and correlate to some extent. Our approach allows for precise, independent manipulation of word frequency and dispersion by inserting an artificial word. By conducting keyness analysis using synthetically created datasets and through inserting an artificial word with precisely manipulated frequency and dispersion into the synthetic dataset, we aim to systematically uncover the characteristics of different measures. Our goal is to determine the degree of sensitivity of each measure to variations in frequency and dispersion. Our method enables us to uncover new advantages and limitations of distinctiveness measures and to compare their sensitivity to frequency and dispersion variations under consistent conditions.

The structure of our paper is as follows: We begin with an overview of previous work in the evaluation of measures of distinctiveness (section 2). Next, we describe our dataset (section 3) and provide a detailed explanation of our methodology (section 4). We then outline our hypotheses (section 5) and present the results of our evaluation (section 6). Finally, we conclude by summarizing our key findings and discussing potential directions for future research (section 7).

2. Previous Work: Evaluation of Keyness Measures

Evaluating measures of distinctiveness is challenging due to the fact that generating a gold standard annotation, based on which performance measures such as precision and recall can be calculated, is very difficult. Distinctiveness is not an inherent characteristic of a word, nor does it depend only on local context; rather, it can only be detected in the context of the entire target corpus while considering it in comparison to another corpus. Therefore, alternative methods of comparison and evaluation of the measures of distinctiveness are required. To tackle this challenge, several studies have attempted to evaluate distinctiveness measures using various methods.

Kilgarriff (2001) examined corpus similarity by reviewing the mathematical characteristics of various distinctiveness measures and argued that the Chi-squared test is the most suitable in finding the most characteristic words of a corpus. Paquot and Bestgen (2009) compared three different measures in their ability to identify frequent and well distributed keywords of academic prose as opposed to fictional prose and discovered that the t-test leads to the best results for their task. Lijffijt et al. (2014) explored a broad

1. We use the term 'synthetic texts' to describe texts that have been generated from documents written by humans through a specific word-level sampling procedure. These texts are therefore different both from 'naturally-occurring' text and from text generated using generative LLMs.

array of measures, focusing on the statistical characteristics of these measures to identify their sensitivity to differences in word frequencies and distributions. The authors randomly sampled a text corpus into two parts in order to minimize differences in both parts and then performed a test for uniformity of p-values. Egbert and Biber (2019) introduced a distinctiveness measure based on dispersion, combining a straightforward dispersion metric with a log-likelihood ratio test. They compare the effectiveness of this approach with corpus frequency methods for identifying distinctive words in online travel blogs. Their study demonstrates that the dispersion-based measure outperforms the other types of measures. Sönning (2023) evaluated 32 metrics, categorized into four dimensions of keyness. Like previously mentioned researchers, he distinguished between two primary perspectives on keyness: frequency-based and dispersion-based measures. His study assessed the effectiveness of these metrics in identifying predefined key verbs in academic writing. The results reveal significant differences among the metrics, with the Wilcoxon rank-sum test and dispersion-based measures emerging as the most effective.

The research we report on here also builds on fundamental work on measures of distinctiveness by our *Zeta and Company* project group. We conducted an in-depth analysis of the qualitative characteristics of these measures (Schröter et al. 2021). To enhance accessibility and usability, we implemented nine measures of distinctiveness in the Python package *pydistinto* (Du et al. 2021b). With Du et al. (2021a), we then introduced a new dispersion-based measure called Eta and compared it with the existing Zeta measure to highlight the advantages and disadvantages of each. Our group also performed a quantitative evaluation of nine measures on natural texts, including several dispersion-based measures, using a downstream classification task (Du et al. 2022). Our approach involved first identifying a given number of distinctive words provided by each measure for novels of a specific genre, in comparison to other literary genres. These distinctive words were then used to classify the novels by genre, with the classification accuracy obtained being a measure of each word list's distinctiveness (in the qualitative sense of discriminatory power). We concluded that dispersion-based measures are more effective than frequency-based measures in identifying characteristic words of a target corpus.

Overall, while previous studies have provided valuable insights into distinctiveness measures, their reliance on abstract statistical analyses, intuitive evaluations, or a narrow selection of measures underscores the need for further research. Our study addresses these limitations by introducing a controlled, synthetic approach with precise manipulation of word frequency and dispersion, while also incorporating a wide range of different measures to enable a more systematic and nuanced assessment of their sensitivity. We have already conducted several analyses using naturally-occurring texts. Now, with our approach using synthetic texts, we aim to test theoretical insights about the measures under specially controlled conditions, allowing for a clearer understanding of how each distinctiveness score is calculated.

We think that using a wide variety of evaluation strategies is most likely to result in robust results, as past experience has shown that even theoretically sound and convincing arguments may not hold up to empirical scrutiny, whether quantitative or qualitative

(as a case in point, consider investigations of distance-based stylometric authorship attribution; Argamon 2007, Evert et al. 2017).

3. Data

Our research is conducted on a synthetic text collection generated through random sampling at the word level from a corpus of French contemporary novels. The foundation for this corpus is a balanced subset from our larger collection of French contemporary popular novels and consists of 320 novels from the 1980s and 1990s. This custom-built corpus maintains equal representation (in terms of the number of novels included), per decade and across four subgroups: literary fiction, sentimental novels, crime fiction novels, and science fiction novels.

The original text corpus comprises approximately 19 million words. We load the entire corpus as a single dataset and randomly sample synthetic 'novels', each with a consistent length of 40,000 words. The sampling was performed at the word level. Our newly generated corpus contains 320 synthetic 'novels', matching the number of novels in the original corpus. This approach addresses two main objectives. First, it ensures that the generated corpus reflects the word occurrences and frequencies as they can be observed in the original corpus. Second, it results in a homogeneous corpus, purposefully eliminating subgenre differences because each text is sampled from the entire corpus.

4. Methods

The objective of our analysis is to assess the hidden properties and limitations of the measures of distinctiveness in identifying distinctive words. This is achieved by applying each measure to a homogeneous synthetic corpus to which an artificial word with a controlled frequency and dispersion has been added. Systematically varying the frequency and dispersion of this word, and observing how its keyness rank in the results varies as a result, shows us to what degree a given keyness measure is sensitive to differences in frequency and/or dispersion. We chose to compare the ranks, rather than the measures' scores, for better comparability.

In our analysis, we have analyzed all nine measures of distinctiveness implemented in our Python package *pydistinto*. The following measures are available in this package: Burrows Zeta (Burrows 2007), logarithmic Zeta (Schöch et al. 2018), Eta (Du et al. 2021a), TF-IDF (Spärck Jones 1972), Wilcoxon rank-sum test (Wilcoxon 1945), Welch's t-test (Welch 1947, Mann and Whitney 1947), the Ratio of relative frequencies (RRF, Gries 2010), the Chi-squared test (Plackett 1983), and the Log-likelihood ratio test (LLR, Dunning 1993).[2] The implemented measures can be categorized into three distinct groups based on their approach to identifying unique keywords when comparing a target and a comparison corpus. Within this framework, the techniques employed can be classified as follows:

1. Frequency-based measures: These measures primarily focus on the frequency of the target word in the corpus, treating the corpus as a 'bag of words' and

2. More information about our rationale for implementing this set of measures in *pydistinto*, as well as detailed descriptions of each measure, can be found in Du et al. 2022.

disregarding how the target word is distributed within the corpus. Examples of measures falling under this classification include the RRF, the Chi-squared test, and the LLR.

2. Distribution-based measures: Rather than just considering corpus-wide mean word frequencies, these measures are based on the distribution of a word (described e.g. via its central tendency and variability) in the corpus. Unlike simpler frequency-based measures, then, these metrics also consider variability indicators, such as standard deviation. They are also quite flexible, in that some of them don't require a normal distribution, allowing for a more nuanced comparison across different distributions. Welch's t-test falls into this category.

3. Dispersion-based measures: These measures evaluate the extent to which the target word is evenly distributed, or dispersed, across a corpus. Measures within this category encompass Burrows Zeta, logarithmic Zeta, Eta, TF-IDF (our implementation of a TF-IDF-based keyness measure), and Wilcoxon rank-sum test (with certain restrictions).[3]

Our approach was as follows: As *pydistinto* requires a certain format of input data (CSV format including the following columns: token, lemma and POS), the original French corpus was annotated with spaCy before randomization.[4] For the analysis with *pydistinto*, we used lemmas as the feature type. At the beginning of the process, the synthetic corpus was divided into segments of equal length, each containing 5,000 words, resulting in 8 segments per novel and a total of 2,560 segments. This segmentation is essential for the calculation of certain measures, such as Zeta and Eta.

Subsequently, the entire corpus was randomly divided into two sub-corpora of equal size for each run of *pydistinto*: target and comparison corpus. An artificial word was then added to both the target and comparison corpus parts with a specified frequency and dispersion.[5] To maintain a constant total word count while adding an artificial word, each instance of the artificial word replaces one instance of an existing word in the corpus.

Our experiment was conducted in two primary settings to investigate the impact of two criteria – the frequency and dispersion of the artificial word within a corpus – on its distinctiveness score, calculated by different measures.

In the first setting, we added an artificial word to only one segment of the target and comparison corpus, albeit with varying frequency. This setting enables us to analyze the influence of only one parameter, namely the frequency. The frequency of the artificial word was set to 10 in the comparison corpus and remained constant there, while varying from 10 to 2,000 words in the target corpus. We used 12 different parameters for the frequency setting in the target corpus (10, 20, 30, 40, 50, 100, 200, 300, 400, 500, 1000, and 2000). For each parameter setting, *pydistinto* was run 100 times to mitigate the impact on the results of high scores for frequent words, which may arise as a result of variation that follows from the random sampling procedure and may in turn influence

3. Note that these latter measures are based on measures of dispersion that are not entirely uncorrelated with frequency (see e.g. Gries 2022). Detailed information about these measures can be found in Du et al. (2022).
4. See: https://spacy.io/ and Honnibal et al. (2020).
5. An artificial word is a specially created combination of letters and numbers that cannot occur in any natural language. An example of an artificial word used in this study looks like the following: untuniutntrng55886.

the distinctiveness score of an artificial word. The corpus was randomly divided into target and comparison parts at the level of the 'novels' for each run. Given the fact that texts were built by randomly sampling words from the entire corpus, and the two subcorpora were built by randomly sampling 'novels' from among all 'novels', any difference between the target and comparison corpora, apart from the artificial word, can only be due to random variation.

In the second setting, we experimented with the dispersion of the artificial word. In this case, the frequency of the artificial word was kept constant at 1,000 occurrences in both the target and comparison corpus, but its dispersion varied in the target corpus while remaining constant in the comparison corpus. The idea was again to isolate one parameter, in this case dispersion, and analyze its influence on the performance of the different measures. For the comparison corpus we used the following settings: we added 1,000 instances of the artificial word to just 1 segment.[6] Dispersion variation was achieved by adding the artificial word with a specified, constant total frequency to the target corpus, but with varying degrees of dispersion. We conducted distinctiveness analyses with variations in the target corpus according to the following schema, where the first number refers to the number of segments that receive the artificial word, and the second to the number of times the artificial word is included in each of the selected segments: 1/1000, 2/500, 5/200, 10/100, 20/50, 50/20, 100/10, 200/5, 500/2, 1000/1. The product of the two values, and therefore the total frequency, remains constant at 1,000 (and is therefore identical to the frequency of the word in the comparison corpus), but the number of segments these occurrences are spread out over is varied systematically. This resulted in a total of 10 parameter settings for the dispersion experiments. Again, *pydistinto* was run 100 times for each parameter setting.

Following this step, the results for each parameter setting were combined into a single dataframe. Subsequently, all words in the corpus were sorted based on their distinctiveness scores, and for each measure, the rank of the artificial word following from its distinctiveness score was recorded. Each measure's performance was evaluated based on the rank of the artificial word (where a rank of 1 indicates the highest distinctiveness score).

5. Hypotheses

For this evaluation experiment, we developed the following hypotheses:

Hypothesis 1. For dispersion-based measures (Eta, Zeta, and logarithmic Zeta, Wilcoxon rank-sum test), we hypothesize that they should not show any variation in scores when frequency changes while dispersion remains constant.

Hypothesis 2. However, dispersion-based measures should be sensitive to even minimal variations in dispersion even when frequency remains constant, as the number of segments containing the target word is crucial for their calculation.

6. In the dispersion analysis, we also tested another scenario, in which we randomly selected 1,000 segments and added one instance of the artificial word to each of them in the comparison corpus, ensuring even dispersion. However, this scenario turned out not to provide significant or additional insights. Therefore, we are not providing further explanations or results here.

Hypothesis 3. We hypothesize that frequency-based measures (RRF, LLR, and chi-square tests) will show high variations in distinctiveness scores even when the frequency difference of an artificial word between the target and comparison corpus is relatively small. This assumption stems from the statistical nature of these measures, which treat a corpus as a bag of words and do not account for word dispersion.

Hypothesis 4. When the frequency of an artificial word is the same in both the target and comparison while its dispersion changes, the scores of frequency-based measures should remain unchanged.

Hypothesis 5. Regarding our TF-IDF-based measure, we expect it to exhibit moderate sensitivity in both frequency and dispersion manipulations. This is because TF-IDF is based on term frequency, but the number of segments containing the target word also significantly influences its calculation.

Hypothesis 6. Regarding Welch's test, we hypothesize that there will be minimal variations in the score in the case of frequency manipulation. This assumption is based on the fact that the calculation of Welch's test relies on the mean and standard deviation of the frequency distributions, rather than on the raw frequency of the word.

6. Results

Because our corpus is based on naturally occurring word frequencies, we conducted an additional analysis to identify potential artifacts caused by random sampling effects in the synthetic texts without the artificial word. This analysis aimed to identify the frequency differences of words in the corpus across multiple runs.

Figure 1 illustrates the relationship between rank and the Ratio of Relative Frequencies (RRF) scores, based on 100 runs of randomly sampled synthetic corpora. As shown, the first rank is typically achieved with RRF scores ranging from 10 to 18. This suggests that, due to the natural variations in the frequencies of existing words, an RRF score below 10 for the artificial word is unlikely to secure the first rank.

As discussed in section 4, we conducted our evaluation in two main settings: frequency variation of an artificial word and dispersion variation. First, we are going to discuss the results of the evaluation based on frequency variations, before moving on to the results for dispersion variations.

6.1 Evaluation Based on Frequency Variations

Concerning the impact of frequency variation on the performance of the measures, as described in section 5, we used 12 different parameters for the frequency settings. Figure 2 depicts the variation in the rank of the artificial word, as calculated by Zeta, Eta, the rank-sum test and Welch's test, respectively, depending on its frequency in the target corpus.[7] The x-axis represents the frequency variation in the target corpus (from 20 to 2,000 occurrences in one segment of the target corpus). On the y-axis, the rank of the artificial word is depicted, each boxplot showing the median and range of the 100 ranks recorded for each particular parameter setting. To enhance the readability of the

7. Zeta and Eta_log are not depicted in the figure, because their results are very similar to the Zeta_log results.

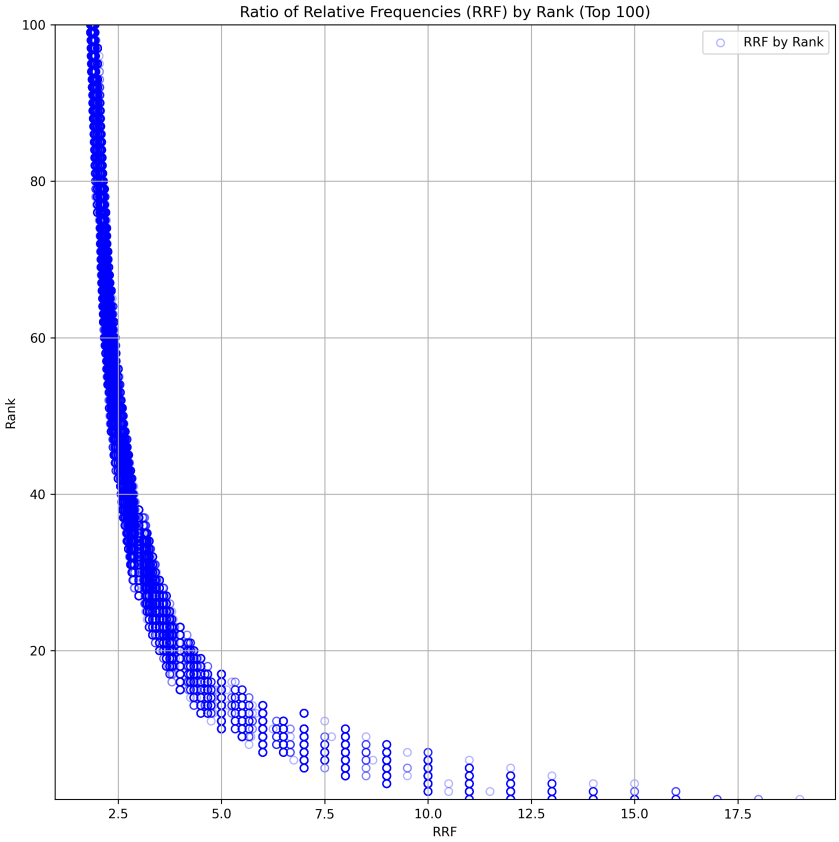

Figure 1: The correlation between the RRF score of the words and their ranks in the synthetic corpus.

figure, the values on the y-axis are presented on a logarithmic scale.

Dispersion-based measures including Zeta, logarithmic Zeta, Eta, Wilcoxon rank-sum test, as well as Welch's t-test, which we consider rather as a distribution-based measure, demonstrate very similar results. For these measures, the frequency variations of an artificial word in the target corpus don't play an important role. The rank of the artificial word consistently exceeds 10,000 for frequencies ranging from 20 to 2,000 in the target corpus, indicating a very low distinctiveness score according to these measures. The scores for Eta, Zeta, logarithmic Zeta, and the Wilcoxon rank-sum tests remain consistent across the board, supporting Hypothesis 1 and validating our method. The scores from Welch's test show minimal variation, as expected in Hypothesis 6.

Frequency-based measures such as the chi-square test, LLR, and RRF exhibit high sensitivity to frequency variations, as expected, supporting Hypothesis 3. However, we can observe some interesting results here. When considering the RRF, the artificial word moves up in rank with increasing frequency from 20 to 100 (Figure 3). Starting from 200 artificial words in the target corpus, RRF-based rank is always 1, which means that the artificial word gets the highest score among all words in the corpus. As for LLR and chi-squared tests, both measures are even more sensitive to frequency variation

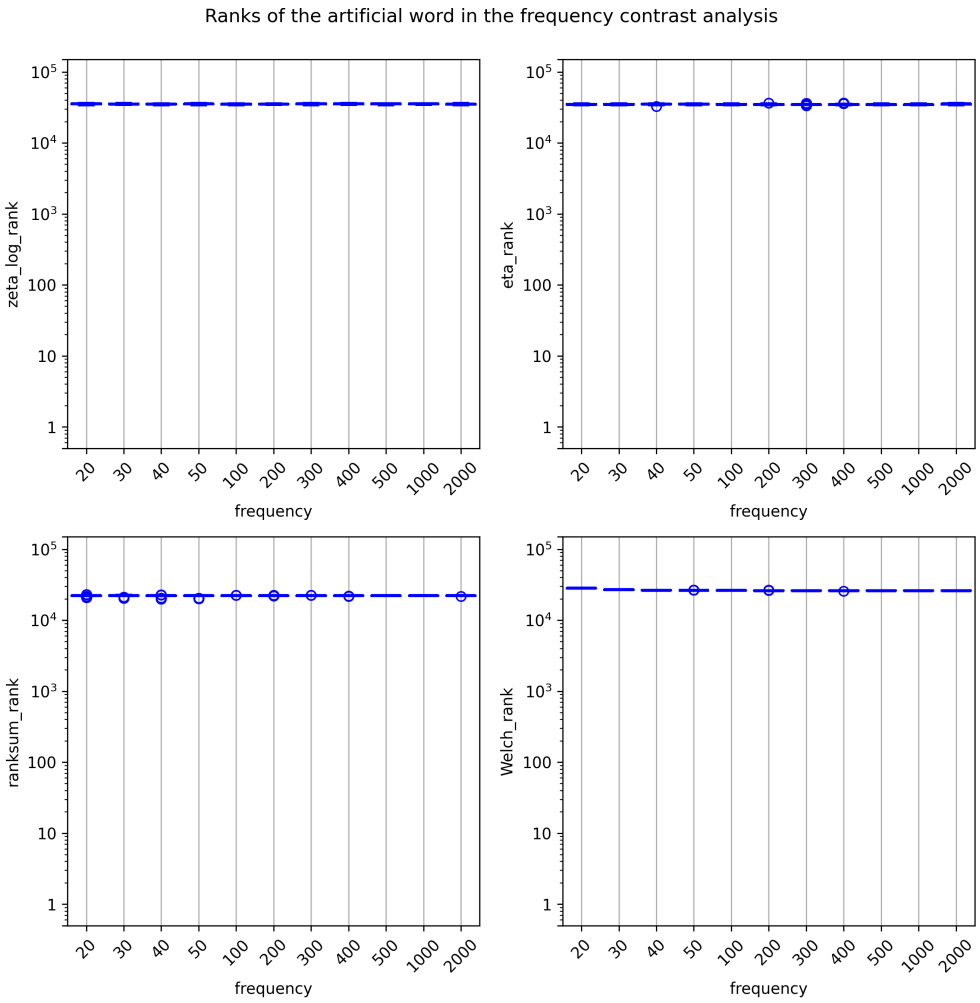

Figure 2: The relation between the frequency of the artificial word in the target corpus and its rank in the results, for Zeta, Eta, rank-sum test and Welch's test.

compared to RRF. Starting at a frequency of just 40, we consistently observe the artificial word achieving the top rank.

TF-IDF is more sensitive to frequency variation than dispersion-based measures but significantly less so than frequency-based measures, aligning with our expectation in Hypothesis 5. With increasing frequency of the artificial word in the target corpus, its rank moves up. Figure 3 shows a moderately strong but continuous rise of the rank of the artificial word.

6.2 Evaluation Based on Dispersion Variations

As previously described, the dispersion analysis was conducted with 1,000 instances of the artificial word in one segment of the comparison corpus. Figure 4 illustrates the variation in the rank of an artificial word calculated by chi-square, LLR, RRF and TF-IDF. The x-axis depicts the dispersion variation of the artificial word in the target corpus from 1/1000 to 1000/1, where the first number represents the number of segments and the second number represents the number of instances of the artificial word distributed over those segments. The dispersion of the artificial word in the comparison corpus remains constant, set at 1/1000, indicating 1,000 words occurring in one segment.

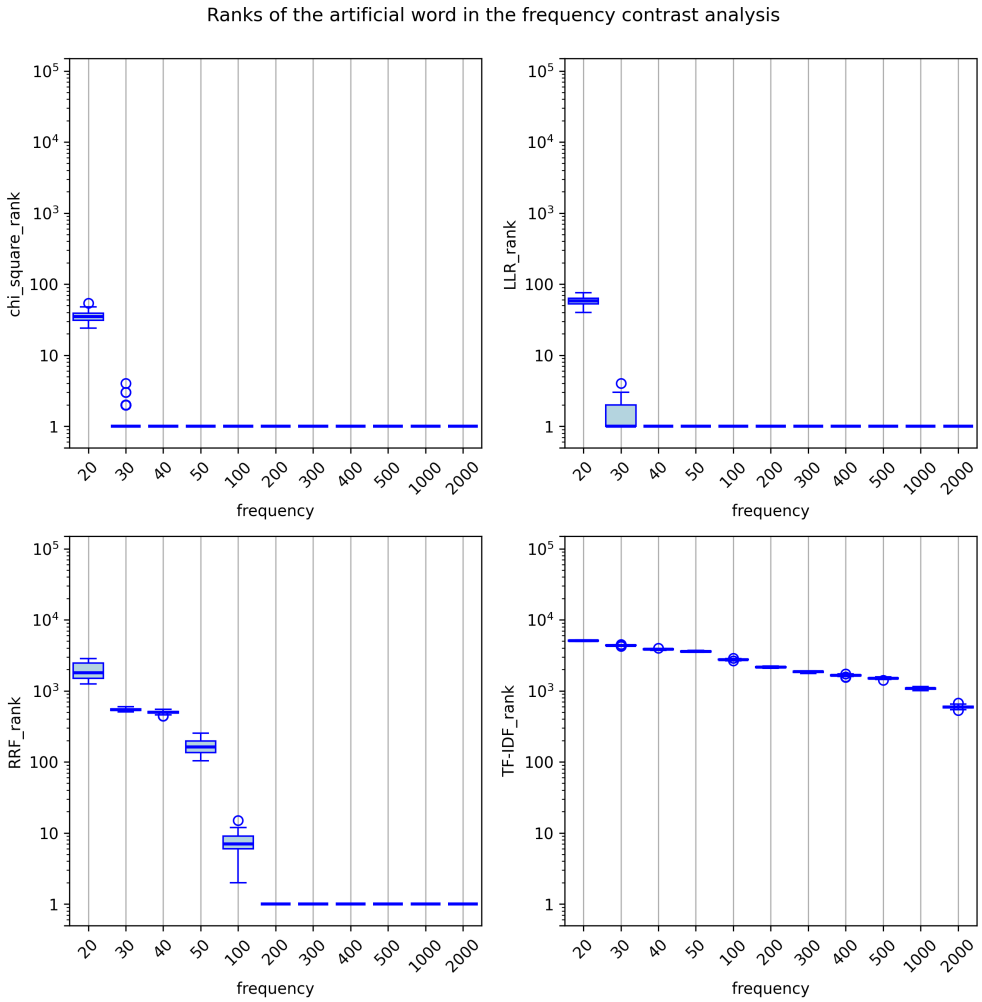

Figure 3: The relation between the frequency of the artificial word in the target corpus and its rank in the results, for RRF, chi-squared test, LLR and TF-IDF.

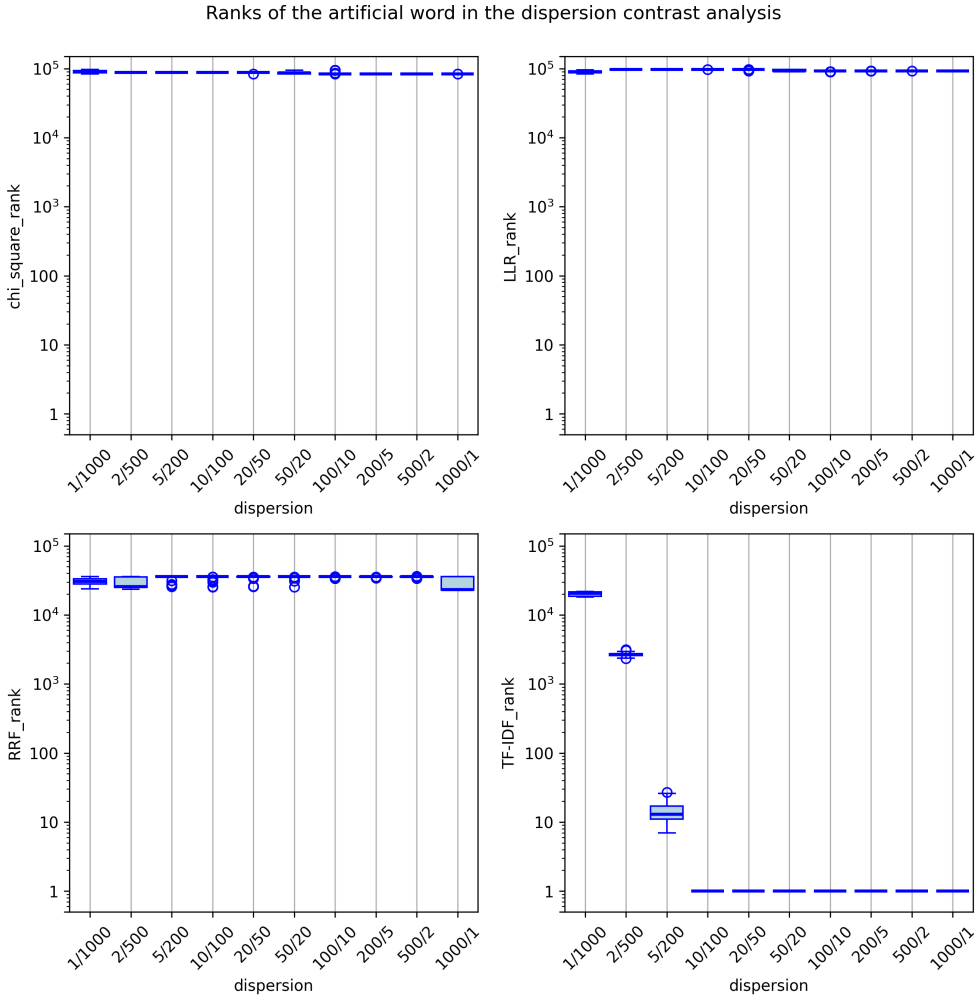

Figure 4: The relation between the dispersion of the artificial word in the target corpus and its rank in the results, for RRF, chi-squared test, LLR and TF-IDF. Dispersion in the comparison corpus is fix at 1/1000.

In these settings, the frequency-based measures produce results consistent with those predicted by Hypothesis 4. When the dispersion changes (while the frequency remains constant), the rank of an artificial word does not change significantly and consistently remains at a level between 10,000 and 100,000.

This is true except for TF-IDF, for which interesting results are observed. Here, we anticipated that, as the dispersion becomes more even, the artificial word would receive a higher score, but only with a moderate rank improvement compared to other dispersion-based measures. In fact, we can observe that TF-IDF scores indeed increase as the number of segments containing the artificial word rises. However, the improvement in scores is not moderate; rather, TF-IDF appears to be highly sensitive to variations in dispersion, which partially rejects Hypothesis 5. We observe the artificial word achieving the top rank starting with a dispersion of just 100 words in 10 segments (Figure 4, bottom right). This oversensitivity implies that the TF-IDF measure fails to distinguish between a dispersion of 100 words across 10 segments vs. one single word across 1,000 segments.

Figure 5 shows results for dispersion variations with the measures Zeta_log, Eta, ranksum and Welch. An interesting result was obtained by Eta. As it is a dispersion-based mea-

sure, we expected Eta to effectively identify an artificial word as distinctive, especially when the word is evenly spread across a high number of segments. However, as the number of segments containing the artificial word in the target corpus increases, its scores remain consistently low compared to randomly assigned words. Only in the most extreme setting, with one occurrence in 1,000 segments in the target corpus, the artificial word receives the top rank (Figure 5).

Regarding the results of Welch's test, when the frequency of the artificial word is identical in both the target and comparison corpora, the score consistently remains zero, resulting in a rank above 10,000. This indicates that, like the frequency-based measures, Welch's test is not sensitive to variations in dispersion within our settings. This actually means that Welch's test is neither sensitive to frequency variations alone, nor to dispersion variations alone.

Regarding the remaining dispersion-based measures, such as both variants of Zeta and the rank-sum test, we observe expected results. With increasing numbers of segments containing the artificial word in the target corpus, the artificial word's rank moves up. Specifically, starting with 10 words in 100 segments, the artificial word consistently receives the top rank according to these three measures (Figure 5). This indicates that Hypothesis 2 is supported solely for these three measures.

7. Conclusion

Conducting analyses of measures of distinctiveness based on synthetic texts, we created ideal conditions to uncover the hidden properties of a range of such measures. Through our experiments, we tested the sensitivity of these measures to variations in the frequency and dispersion of a specific word. In many cases, our hypotheses regarding the performance of the measures were confirmed. Frequency-based measures are not sensitive to variations in dispersion, while dispersion-based measures are not affected by frequency variations. These observations are not surprising, of course, but they do validate our method.

However, some hypotheses were partly rejected and we have also uncovered some previously unknown (or at least undocumented) properties of measures of distinctiveness. In particular, we found that LLR and chi-squared tests are even more sensitive to frequency variation than RRF. For this reason, we generally do not recommend using the LLR and chi-squared tests, as they are highly sensitive to changes in frequency and are therefore not well-suited for keyness analysis aimed at identifying important content words. Both Zeta variations and the rank-sum test demonstrated similar scores and abilities to detect distinctive words, including cases in which the differences only concern the dispersion of words. Moreover, we discovered that TF-IDF is highly sensitive to dispersion differences of the target word, compared to other dispersion-based measures. Finally, we found that Eta cannot detect a word with a clear contrast in dispersion when its frequency is the same in both the target and comparison corpora. In our evaluation we observed words steadily moving up in rank with Zeta and rank-sum, while TF-IDF and Eta show more abrupt increases. We suggest that a gradual, continuous rank improvement is a desirable characteristic of a distinctiveness measure, as it indicates better sensitivity to slight variations in dispersion and is likely to produce more predictable results. For

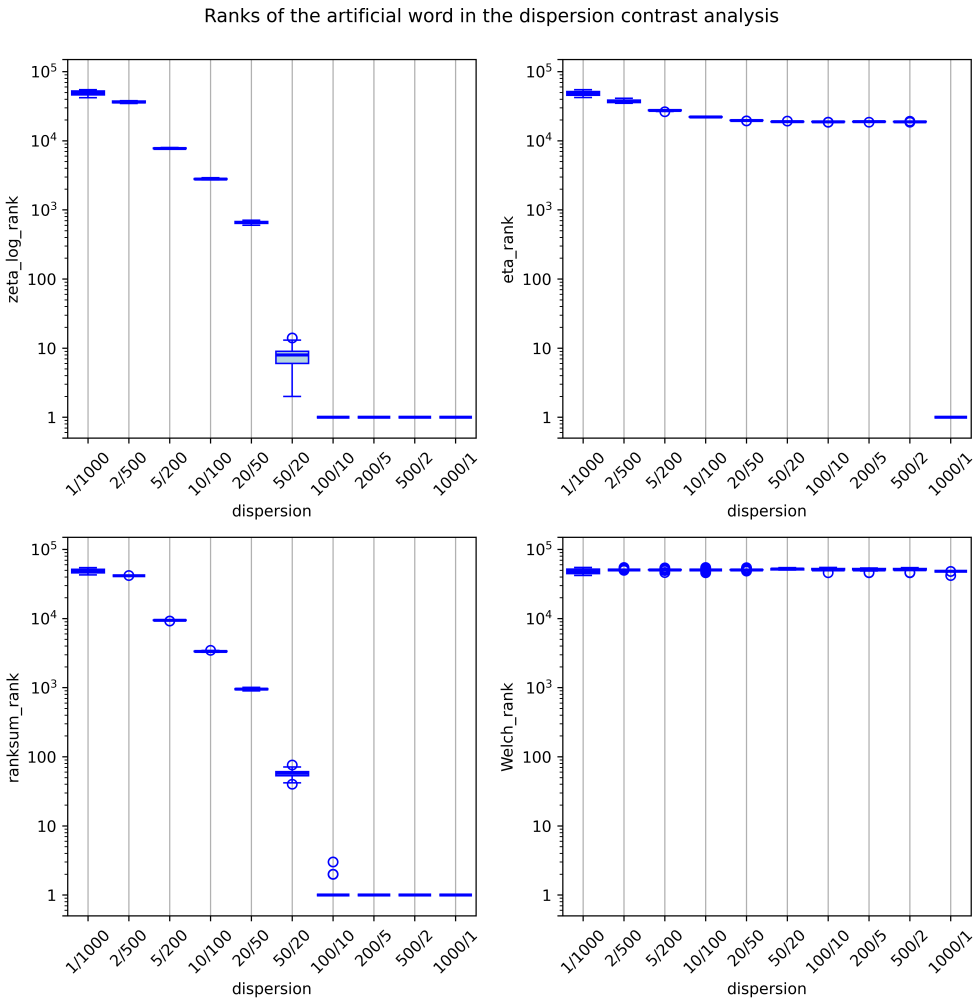

Figure 5: The relation between the dispersion of the artificial word in the target corpus and its rank in the results, for Zeta, Eta, rank-sum test and Welch's test. Dispersion in the comparison corpus is constant at 1/1000.

example, if a researcher is interested in identifying words that display contrasting dispersion within two subcorpora, without considering their frequency, then Zeta and the rank-sum test would be most appropriate for this task.

Despite the interesting observations derived from these analyses, there is significant potential for future work. One key step is to extend our framework by implementing additional measures of distinctiveness. Another area for future work involves expanding our analysis by implementing additional parameter settings that combine frequency and dispersion variations of the artificial word. Isolating dispersion or frequency often results in constant scores from the measures, but combining these parameters promises to provide new opportunities to uncover additional properties of these measures. A final, crucial step is to explore practical applications of this newfound knowledge about distinctiveness measures. Understanding the specific contexts and scenarios in which each of these measures can be most effectively utilized will open up new possibilities and enhance our ability to analyze and compare textual corpora more accurately.

8. Data Availability

Data can be found here: https://github.com/Zeta-and-Company/synthetic_texts_evaluation. It has been archived and is persistently available at https://doi.org/10.5281/zenodo.15525428.

9. Software Availability

Software can be found here: https://github.com/Zeta-and-Company/synthetic_texts_evaluation. It has been archived and is persistently available at https://doi.org/10.5281/zenodo.15525428.

10. Author Contributions

Julia Havrylash: Conceptualization, Data Curation, Methodology, Formal Analysis, Software, Visualisation, Writing – original draft, Writing – review & editing

Christof Schöch: Funding Acquisition, Supervision, Visualisation, Writing – review & editing

References

Argamon, Shlomo (2007). "Interpreting Burrows's Delta: Geometric and Probabilistic Foundations". In: *Literary and Linguistic Computing* 23 (2), 131–147. 10.1093/llc/fqn003.

Burrows, John (2007). "All the Way Through: Testing for Authorship in Different Frequency Strata". In: *Literary and Linguistic Computing* 22 (1), 27–47. 10.1093/llc/fqi067.

Culpeper, Jonathan (2009). "Keyness: Words, Parts-of-Speech and Semantic Categories in the Character-Talk of Shakespeare's *Romeo and Juliet*". In: *International Journal of Corpus Linguistics* 14 (1), 29–59. 10.1075/ijcl.14.1.03cul.

Du, Keli, Julia Dudar, Cora Rok, and Christof Schöch (2021a). "Zeta & Eta: An Exploration and Evaluation of Two Dispersion-based Measures of Distinctiveness". In: *Proceedings of Computational Humanities Research 2021*. Ed. by Maud Ehrmann, Folgert Karsdorp, Melvin Wevers, Tara Lee Andrews, Manuel Burghardt, Mike Kestemont, Enrique Manjavacas, Michael Piotrowski, and Joris van Zundert, 181–194. http://ceur-ws.org/Vol-2989/short_paper11.pdf (visited on 10/14/2025).

Du, Keli, Julia Dudar, and Christof Schöch (2021b). *Pydistinto - a Python Implementation of Different Measures of Distinctiveness for Contrastive Text Analysis*. Version v0.1.1. Zenodo. 10.5281/zenodo.5245096.

— (2022). "Evaluation of Measures of Distinctiveness: Classification of Literary Texts on the Basis of Distinctive Words". In: *Journal of Computational Literary Studies* 1 (1). 10.48694/jcls.102.

Dunning, Ted (1993). "Accurate Methods for the Statistics of Surprise and Coincidence". In: *Computational Linguistics* 19 (1), 61–74. http://aclweb.org/anthology/J93-1003 (visited on 10/16/2025).

Egbert, Jesse and Doug Biber (2019). "Incorporating Text Dispersion into Keyword Analyses". In: *Corpora* 14 (1), 77–104. 10.3366/cor.2019.0162.

Evert, Stefan, Fotis Jannidis, Thomas Proisl, Steffen Pielström, Thorsten Vitt, Christof Schöch, and Isabella Reger (2017). "Understanding and Explaining Distance Measures for Authorship Attribution". In: *Digital Scholarship in the Humanities* 23 (suppl_2). 10.1093/llc/fqx023.

Gonon, Laetitia, Vannina Goossens, Olivier Kraif, Iva Novakova, and Julie Sorba (2018). "Motifs Textuels Spécifiques Au Genre Policier et à La Littérature Blanche". In: *6e Congrès Mondial de Linguistique Française, SHS Web of Conferences* 46. Ed. by Franck Neveu, Bernard Harmegnies, Linda Hriba, and Sophie Prévost. 10.1051/shsconf/20184606007.

Gries, Stefan Th. (2010). "Useful Statistics for Corpus Linguistics". In: *A Mosaic of Corpus Linguistics: Selected Approaches*. Ed. by Aquilino Sánchez and Moisés Almela. Peter Lang, 269–291.

— (2022). "What Do (Most of) Our Dispersion Measures Measure (Most)? Dispersion?" In: *Journal of Second Language Studies* 5 (2), 171–205. 10.1075/jsls.21029.gri.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd (2020). "spaCy: Industrial-strength Natural Language Processing in Python". In: *Zenodo*. 10.5281/zenodo.1212303.

Kilgarriff, Adam (2001). "Comparing Corpora". In: *International Journal of Corpus Linguistics* 6 (1), 97–133. 10.1075/ijcl.6.1.05kil.

Lijffijt, Jefrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila (2014). "Significance Testing of Word Frequencies in Corpora". In: *Digital Scholarship in the Humanities* 31 (2), 374–397. 10.1093/llc/fqu064.

Mann, H. B. and D. R. Whitney (1947). "On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other". In: *The Annals of Mathematical Statistics* 18 (1), 50–60. 10.1214/aoms/1177730491.

Paquot, Magali and Yves Bestgen (2009). "Distinctive Words in Academic Writing: A Comparison of Three Statistical Tests for Keyword Extraction". In: *Corpora: Pragmatics and Discourse*. Ed. by Andreas H. Jucker, Daniel Schreier, and Marianne Hundt. Brill | Rodopi, 247–269. 10.1163/9789042029101_014.

Plackett, Robin L. (1983). "Karl Pearson and the Chi-Squared Test". In: *International Statistical Review / Revue Internationale de Statistique* 51 (1). 10.2307/1402731.

Schöch, Christof (2018). "Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie". In: *Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven*. Ed. by Toni Bernhart, Sandra Richter, Marcus Lepper, Marcus Willand, and Andrea Albrecht. De Gruyter, 77–94. 10.1515/9783110523300-004.

Schöch, Christof, Daniel Schlör, Albin Zehe, Henning Gebhard, Martin Becker, and Andreas Hotho (2018). "Burrows' Zeta: Exploring and Evaluating Variants and-Parameters". In: *Book of Abstracts of the Digital Humanities Conference 2018*. Ed. by Jonathan Girón Palau and Isabel Galina Russell. ADHO. https://dh2018.adho.org/en/burrows-zeta-exploring-and-evaluating-variants-and-parameters/ (visited on 10/16/2025).

Schröter, Julian, Keli Du, Julia Dudar, Cora Rok, and Christof Schöch (2021). "From Keyness to Distinctiveness – Triangulation and Evaluation in Computational Literary Studies". In: *Journal of Literary Theory* 15 (1-2), 81–108. 10.1515/jlt-2021-2011.

Sönning, Lukas (2023). "Evaluation of Keyness Metrics: Performance and Reliability". In: *Corpus Linguistics and Linguistic Theory* 20 (2), 263–288. 10.1515/cllt-2022-0116.

Spärck Jones, Karen (1972). "A Statistical Interpretation of Term Specificity and Its Application in Retrieval". In: *Journal of Documentation* 28 (1), 11–21. https://dl.acm.org/doi/10.5555/106765.106782 (visited on 10/14/2025).

Weidman, Sean G. and James O'Sullivan (2018). "The Limits of Distinctive Words: Re-evaluating Literature's Gender Marker Debate". In: *Digital Scholarship in the Humanities* 33 (2), 374–390. 10.1093/llc/fqx017.

Welch, Bernard Lewis (1947). "The Generalization of Student's Problem When Several Different Population Variances Are Involved". In: *Biometrika* 34 (1-2), 28–35. 10.1093/biomet/34.1-2.28.

Wilcoxon, Frank (1945). "Individual Comparisons by Ranking Methods". In: *Biometrics Bulletin* 1 (6). 10.2307/3001968.