Exploring Measures of Distinctiveness: An Evaluation Using Synthetic Texts

Dudar, Julia
dudar@uni-trier.de
Universität Trier, Deutschland
ORCID: 0000-0001-5545-9562

Schöch, Christof
schoech@uni-trier.de
Universität Trier, Deutschland
ORCID: 0000-0002-4557-2753

Introduction

Comparing groups of texts with each other in order to investigate what is characteristic about each group is a fundamental approach used in many research contexts, and measures of distinctiveness (also known as keyness measures) support such research in a quantitative perspective. The research we report on here is a further step in our fundamental work on measures of distinctiveness used for comparison of groups of texts.

In the research reported on here, we focus on evaluating measures of distinctiveness through an analysis based on synthetic texts. Recent work has shown the importance of both frequency and dispersion of words for keyness analysis (Gries, 2021). 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 sensitivity of each measure to variations in frequency and dispersion.

Evaluating measures of distinctiveness

Evaluating measures of distinctiveness is challenging due to the fact that generating a gold-standard annotation is not possible. Distinctiveness is not an inherent characteristic of a word but can only be detected in the context of the entire target corpus and in comparison to another corpus. 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. Paquot & Bestgen (2009) compared three different measures in their ability to identify words distinctive of academic prose as opposed to fictional prose. Lijffijt et al. (2014) explored a broad array of measures, focusing on the statistical characteristics of these measures.

Within the framework of our project "Zeta and Company," 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., 2021a). Subsequently, we performed an evaluation of two dispersion-based measures (Du et al., 2021b) and a quantitative evaluation of nine measures (Du et al., 2022).

Our research proposes a new method for evaluating measures of distinctiveness, utilizing synthetically created text collections that reflect word frequencies as they occur in a real corpus, but within an artificially homogeneous corpus design. Studies based on naturally-occurring language must work around the fact that frequency and dispersion of any word will both vary and correlate to some extent. Our approach allows for precise, independent manipulation of word frequency and dispersion by inserting an artificial word. 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.

Data

Our research is conducted on a synthetic text collection generated through random sampling 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 first published during the time period 1980 to 1999. This custom-built corpus maintains equal representation, per decade, across four subgroups: highbrow novels and lowbrow novels with three subgenres (sentimental novels, crime fiction, and science fiction).

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. Our newly-generated corpus contains 320 synthetic "novels," matching the number of novels in the original corpus. This approach addresses two main objectives. Firstly, it ensures that the generated corpus reflects real-life word occurrences and frequencies. Secondly, it results in a homogeneous corpus, eliminating subgenre differences, since each text is sampled from the entire corpus.

Method

To conduct our evaluation we used nine measures of distinctiveness implemented in our Python package *pydistinto*.

They can be classified into three groups: frequency-based, distribution-based, and primarily dispersion-based measures. Frequency-based measures, including the Ratio of relative frequencies (RRF), chi-square test, and Log-likelihood ratio test (LLR), focus on word frequency, ignoring distribution and dispersion. Distribution-based measures, like Welch's t-test, assess deviations from the central tendency within each text group. Dispersion-based measures, including Burrows Zeta, logarithmic Zeta, Eta, our implementation of TF-IDF-based measure, and Wilcoxon rank-sum test, evaluate how evenly a word is distributed across each text group.

The original French corpus was annotated with spaCy to create the input required by *pydistinto*, and lemmas were used as feature types. This annotation structure was also retained for each word in the synthetic corpus. During the analysis, the synthetic corpus was segmented into 2560 segments of 5000 words each.

Our experiment had two primary settings to assess the impact of frequency and dispersion on distinctiveness scores.

In the first setting, an artificial word was added to one segment of both the target and comparison corpus. This setting enables us to analyze the influence of only one parameter, namely the frequency. To maintain a constant total word count while adding an artificial word, other words in the corpus that occupied the same position as the artificial word were replaced. The frequency of the artificial word was constant (10 words) in the comparison corpus, but varied in the target corpus (10 to 2000 words). *Pydistinto* was run 100 times for each of 12 frequency settings, to mitigate the occasional impact on the results of high scores for frequent words. The corpus was randomly divided into target and comparison parts for each run. Given the fact that texts were built by randomly sampling words from the entire corpus, any difference between the target and comparison corpora, apart from the artificial word, can only be due to random variation.

In the second setting, the artificial word's frequency was fixed at 1000 occurrences, but its dispersion varied. The idea was again to isolate one parameter, in this case dispersion, and analyze its influence on the performance of measures. The dispersion experiment involved different numbers of segments receiving the artificial word with specified frequencies, leading to 20 parameter settings. In the comparison corpus, scenarios included 1 segment with 1000 words and 1000 segments with 1 word. For each of these parameters, we conducted distinctiveness analyses with variations in the target corpus (the first number refers to the number of segments that receive the artificial word, and the second to how many 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. *Pydistinto* was again run 100 times for each parameter setting.

The results were compiled into a single dataframe, with all words in the corpus sorted and ranked by their distinctiveness scores. Each measure's performance was evaluated based on the rank of the artificial word (where a rank of 1 indicates the highest distinctiveness score).

Hypotheses

1. 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.
2. 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 will remain unchanged.
3. 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.
4. However, they should be sensitive to variations in dispersion when frequency remains constant, as the number of segments containing the target word is crucial for their calculation.
5. Regarding 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.

Results

Since our corpus is based on naturally occurring word frequencies, we conducted an additional analysis to evaluate the 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 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.

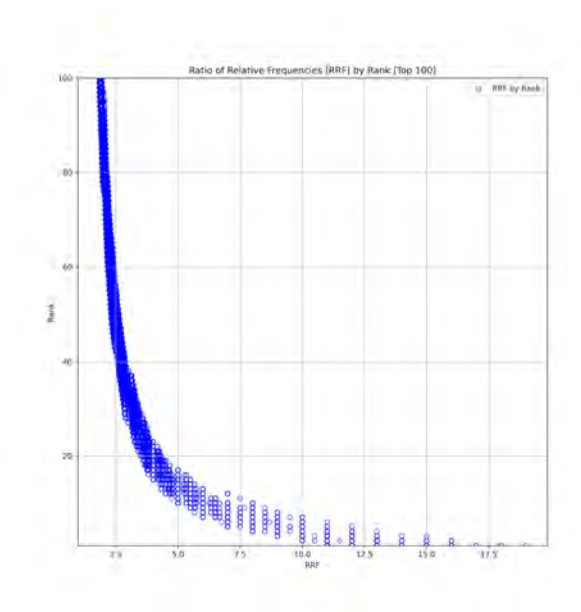

Figure 1. The correlation between the RRF score of the words and their ranks in the synthetic corpus.

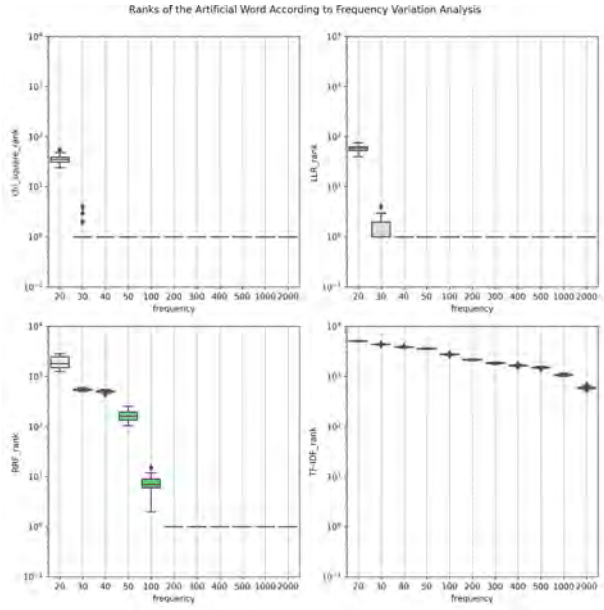

Figure 2. The relationship between the frequency of the artificial word in the target corpus and its rank in the results for RRF, chi-squared test, LLR and TF-IDF. Frequency in the comparison corpus is constant at 10.

In describing the results, our main focus lies on unexpected interesting observations, rather than on a description of all findings. Generally speaking, our expectations formulated in Hypotheses 1-4 are confirmed: Frequency-based measures are sensitive to differences in frequency but not to dispersion, and dispersion-based measures are sensitive to differences in dispersion, but not in frequency. However, this result comes with many nuances.

Analysis based on frequency variations

Analyzing the frequency-based measures such as the chi-square test, LLR, and RRF, we observe a tendency for the score to increase with increasing frequency, but there are notable differences among the various measures (Fig. 2). When the artificial word reaches a frequency of 200 or higher in the target corpus, its RRF rank is consistently 1, indicating that it achieves the highest score among all words in the corpus. This observation aligns with our earlier analysis conducted without the artificial word (Fig. 1). Notably, RRF scores of 10 or below (corresponding to a frequency of 100 words in the target corpus) fail to achieve the first rank. This also explains the wide distribution of ranks observed for RRF scores based on a frequency of 100 words in the target corpus. Regarding the LLR and chi-squared tests, both measures are even more sensitive to frequency variation compared to RRF. At a frequency of 40 and higher, we observe the artificial word achieving the first rank. TF-IDF shows moderate sensitivity to frequency variation, partially supporting Hypothesis 5.

Analysis based on dispersion variations

Regarding the performance of dispersion-based measures, such as both variants of Zeta and the rank-sum test, when the artificial word is inserted into only one segment of the comparison corpus and the number of segments containing the artificial word in the target corpus increases, the word moves up in the ranking. Specifically, starting with 10 words in 100 segments, the artificial word almost consistently receives the first rank according to these three measures (Fig. 3).

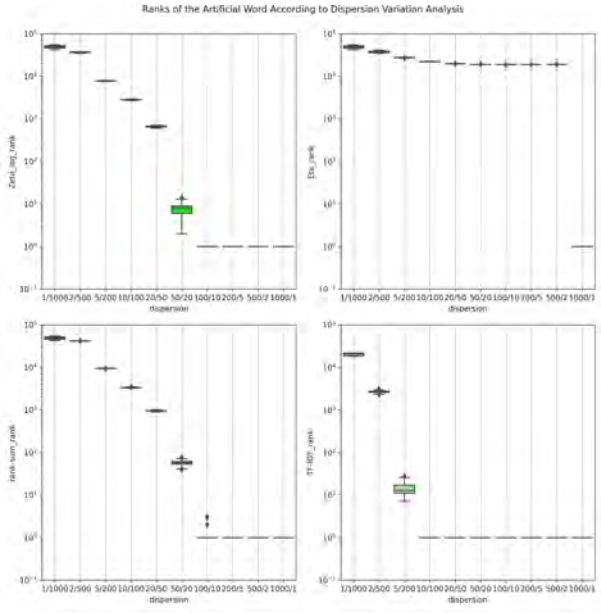

Figure 3. The relationship between the dispersion of the artificial word in the target corpus and its rank in the results for logarithmic Zeta, rank-sum test, TF-IDF and Eta. Dispersion in the comparison corpus is constant at 1/1000.

Eta shows interesting results in these settings. As a dispersion-based measure, 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 increases, its scores remain consistently low compared to randomly assigned words. Only in a scenario with one occurrence in 1000 segments does the artificial word receive the first rank (Fig. 3). This indicates that Hypothesis 4 is supported solely by both variants of Zeta and the rank-sum test.

Unexpected results are also observed with TF-IDF. Similarly to the results of dispersion-based measures, as the number of segments increases, the rank of the artificial word moves up. However, in contrast to the moderate movement with respect to rank seen with dispersion-based measures, in the scenario with 1000 words in one segment of the comparison corpus, the artificial word achieves the first rank starting with a dispersion of just 100 words in 10 segments (Fig. 3). This result partially contradicts our expectation in Hypothesis 5 regarding the moderate sensitivity of TF-IDF to variations in dispersion.

Conclusion

Conducting analyses based on synthetic texts, we created ideal conditions to uncover the hidden properties of a range of distinctiveness measures. Through our experiment, we tested the sensitivity of these measures to variations in the frequency and dispersion of a specific word. We found that LLR and chi-square tests are even more sensitive to frequency variation than RRF, which is simple and relies only on word frequency. Both Zeta variations and the rank-sum test demonstrated similar scores and abilities to detect distinctive words. Moreover, we discovered that TF-IDF is more sensitive to slight dispersion differences of the target word compared to other dispersion-based measures. Finally, we found that Eta does not detect a word with a clear contrast in dispersion when its frequency is the same in both the target and comparison corpora.

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 new measures of distinctiveness, particularly those that rely purely on dispersion rather than doing so only primarily, in a combination of frequency and dispersion. Another crucial step is to explore practical applications of this newfound knowledge about distinctiveness measures. Understanding the specific contexts and scenarios where these measures can be effectively utilized will open up new possibilities and enhance our ability to analyze and compare textual corpora more predictably and more accurately.

Additionally, this approach can easily be applied to corpora in languages other than French. While we assume that the method will work similarly with other languages, we encourage other researchers to test our method on further corpora to validate the robustness of our results.

Data and Code

Data and Code Repository: https://github.com/Zeta-and-Company/synthetic_texts_evaluation.

Bibliographie

Gries, Stefan Th. 2021. A new approach to (key) keywords analysis: Using frequency, and now also dispersion. Research in Corpus Linguistics, 9(2), 1-33. https://doi.org/10.32714/ricl.09.02.02 .

Du, Keli, Julia Dudar, and Christof Schöch. 2021a. ‘Pydistinto - a Python Implementation of Different Measures of Distinctiveness for Contrastive Text Analysis’. Zenodo. https://doi.org/10.5281/ZENODO.5245096 .

Du, Keli, Julia Dudar, Cora Rok, and Christof Schöch. 2021b. ‘Zeta & Eta: An Exploration and Evaluation of Two Dispersion-Based Measures of Distinctiveness’. Proceedings Computational Humanities Research 2021. http://ceur-ws.org/Vol-2989/short_paper11.pdf .

Du, Keli, Julia Dudar, and Christof Schöch. 2022. ‘Evaluation of Measures of Distinctiveness: Classification of Literary Texts on the Basis of Distinctive Words’. Journal of Computational Literary Studies 1 (1). https://doi.org/10.48694/JCLS.102 .

Kilgarriff, Adam. 2001. ‘Comparing Corpora’. International Journal of Corpus Linguistics 6 (1): 97–133. https://doi.org/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'. Digital Scholarship in the Humanities 31 (2): 374–97. https://doi.org/10.1093/llc/fqu064 .

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, edited by Andreas H. Jucker, Daniel Schreier, and Marianne Hundt. Brill | Rodopi. https://doi.org/10.1163/9789042029101_014 .

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