jcls. Journal of Computational Literary Studies

OPEN ACCESS

Article

Evaluation of Measures of Distinctiveness

Classification of Literary Texts on the Basis of Distinctive Words

Keli Du[1]
Julia Dudar[2]
Christof Schöch[2]

1. Trier Center for Digital Humanities, University of Trier, Trier, Germany.
2. Department for Computational Linguistics and Digital Humanities, University of Trier, Trier, Germany.

Citation
Keli Du, Julia Dudar, and Christof Schöch (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

Date published 2022-11-30
Date accepted 2022-04-21
Date received 2021-12-22

Keywords
distinctiveness, keyness, evaluation, literary texts

Reviewers
Melanie Andresen
Daniil Skorinkin

Note
This paper has passed through the conference track of JCLS. In addition to being peer reviewed, it was presented and discussed at the 1st Annual Conference of Computational Literary Studies 2022 at the Technical University of Darmstadt.

Abstract. This paper concerns an empirical evaluation of nine different measures of distinctiveness or 'keyness' in the context of Computational Literary Studies. We use nine different sets of literary texts (specifically, novels) written in seven different languages as a basis for this evaluation. The evaluation is performed as a downstream classification task, where segments of the novels need to be classified by subgenre or period of first publication. The classifier receives different numbers of features identified using different measures of distinctiveness. The main contribution of our paper is that we can show that across a wide variety of parameters, but especially when only a small number of features is used, (more recent) dispersion-based measures very often outperform other (more established) frequency-based measures by significant margins. Our findings support an emerging trend to consider dispersion as an important property of words in addition to frequency.

1. Introduction

Edward Tufte, the pioneer of data visualization, famously wrote: "At the heart of quantitative reasoning is a single question: Compared to what?" (Tufte 1990, 67). And indeed, any number or value established in some way can only really be endowed with meaning when it is placed in the context of other, comparable numbers or values. One may think of several fundamental strategies for such a contextualization of numbers. Taking the same measurement at different times is one such strategy and taking the same measurement in different subsets of a dataset is another. Each of these strategies comes with typical statistical operations for the comparison of the values, such as regression to determine a trend over time or a test of statistical significance to compare the distributions of values in two subsets of a dataset (Diez et al. 2019).

What the above observation points to is that comparison is a fundamental operation in many domains operating with numerical values. This is also true, however, for many text-based domains of research, whether statistically-oriented or not (Klimek and Müller 2015). The research we report on here brings both strands together in the sense that it is located at the intersection of literary studies and statistics. More precisely, our research is concerned with modeling, implementing, evaluating and using statistical measures of comparison of two or several groups of texts. The measures we focus on are used to identify characteristic or distinctive features of each group of texts in order to gain an evidence-based understanding of the specific contents, style and/or structure of these groups of texts. As we describe below, such measures have been developed in domains such as Information Retrieval (IR), Corpus and Computational Linguistics (CL), or Computational Literary Studies (CLS). In our research, we bring together knowledge and insight from these domains with the general objective of fostering a better understanding of measures of distinctiveness.

The research we report on in this contribution is set in the wider context of our research into measures of distinctiveness for comparison of groups of texts. Previously, we have worked on the issue of qualitative validation of measures of distinctiveness (see Schröter et al. 2021). We have also implemented a wide range of measures of distinctiveness in our Python package *pydistinto*.[1] With the current contribution, we focus on the step of evaluating the performance of a substantial range of such measures using a downstream classification task.

In this paper, we focus mainly on subgenres of the novel as our dinstinguishing category. This is motivated both by the fact that subgenres are an important classificatory principle in Literary Studies[2] and by our anecdotal observation that human readers of popular literature are able to determine the subgenre of a novel (whether they are reading a crime fiction, sentimental, or science-fiction novel) based on only a relatively small section from a given novel. The classification task we use in this contribution is meant to mirror this ability and asks the following question: How reliably can a machine learning classifier, based on words identified using a given measure of distinctiveness, identify the subgenre of a novel when provided only with a short segment of that novel? The subgenre labels used in this task are derived from publisher data, especially with respect to book series dedicated to specific subgenres of the novel. We test the identification of distinctive words with a wide range of measures of distinctiveness (including measures that can be described as frequency-based, distribution-based, and dispersion-based) and using a broad range of literary corpora in seven different languages.

Specifically for the task at hand, we further hypothesize that dispersion-based measures of distinctiveness should have an advantage over other measures. The reason for this, we assume, is twofold: first, features (single word forms, in our case) identified to be distinctive by a dispersion-based measure have a higher chance of appearing in shorter, randomly selected segments taken from an entire novel than features identified using other kinds of measures, in particular frequency-based measures; second, dispersion-based measures have a tendency to identify content-related words as distinctive, in contrast to (some) frequency-based measures, which tend to identify high-frequency function words as distinctive (as observed in Schöch et al. 2018).

Our paper is structured as follows: First, we summarize related work (a) describing different measures of distinctiveness and (b) specifically comparing several measures of distinctiveness to each other (Section 2). We go on to describe the different corpora

[1] See: `https://github.com/Zeta-and-Company/pydistinto` (DOI: `https://doi.org/10.5281/zenodo.6517683`).
[2] For a concise introduction to genre theory, see Hempfer (2014) and, with a focus on computational approaches to genre, Schöch (2022).

we have used for our study (Section 3) as well as the methods used to perform the evaluation task and to analyze the results (Section 4). We then discuss the results we have obtained, first in a single-language setting, then in a multi-language setting (Section 5). We close our contribution by summarizing our key findings and describing possible future work (Section 6).

2. Related Work

Related work falls into two groups, either defining and/or describing one or several measures of 'keyness' or distinctiveness, or specifically comparing several measures of distinctiveness to each other based on their mathematical properties or on their performance.

2.1 Measures of Distinctiveness

The measures of distinctiveness implemented in our framework have their origins in the disciplines of IR, CL, and CLS.

Name	Type of measure	References	Evaluated in
TF-IDF	Term weighting	Luhn 1957; Spärck Jones 1972	Salton and Buckley 1988
Ratio of relative frequencies (RRF)	Frequency-based	Damerau 1993	Gries 2010
Chi-squared test (χ^2)	Frequency-based	Dunning 1993	Lijffijt et al. 2014
Log-likelihood ratio test (LLR)	Frequency-based	Dunning 1993	Egbert and Biber 2019; Paquot and Bestgen 2009; Lijffijt et al. 2014
Welch's t-test (Welch)	Distribution-based	Welch 1947	Paquot and Bestgen 2009 (t-test); Lijffijt et al. 2014
Wilcoxon rank sum test (Wilcoxon)	Dispersion-based	Wilcoxon 1945; Mann and Whitney 1947	Paquot and Bestgen 2009; Lijffijt et al. 2014
Burrows Zeta (Zeta_orig)	Dispersion-based	Burrows 2007; Craig and Kinney 2009	Schöch 2018
logarithmic Zeta (Zeta_log)	Dispersion-based	Schöch 2018	Schöch 2018; Du et al. 2021
Eta	Dispersion-based	Du et al. 2021	Du et al. 2021

Table 1: An overview of measures the of distinctiveness

Table 1 gives a short overview of the measures of distinctiveness implemented in our Python library, along with their references and information about studies in which they were evaluated. Under the heading 'Type of measure', we very roughly characterize the underlying kind of quantification of the unit of measurement. As all the measures have different mathematical calculations and describing all of them in detail goes beyond the scope of this paper, we propose this typology as a brief and simplified review that summarizes the key characteristics of the implemented measures.

In **Information Retrieval**, identifying distinctive features of given documents is a fundamental and necessary task when it comes to extracting relevant documents for specific terms, keywords or queries. The most widespread ‘keyness’ measure in this domain is the term frequency - inverse document frequency measure (TF-IDF). It was first suggested by Luhn (1957) and optimized by Spärck Jones (1972). It weighs how important a word is to a document in a collection of texts. Today, there is a wide range of different variants and applications of the TF-IDF measure. One prominent example is the TF-IDF-Vectorizer contained in the Python library *scikit-learn* that suggests many useful parameters. The TF-IDF measure implemented in our framework is based on this library.

When it comes to the amount and the variety of measures of distinctiveness, **Computational Linguistics** is the most productive domain. However, almost all measures widely used in CL were originally not invented for text analysis, but were adapted from statistics. As they are usually used in CL for corpus analysis, many of them are implemented in different corpus analysis tools.

One of the simplest measures is the ratio of relative frequencies (Damerau 1993). As its name already says, it considers only the relative frequency of features and relies on the division of the value for the target corpus by the value of the comparison corpus. It cannot deal with words that do not appear in the comparison corpus.

The Chi-squared (χ^2) and log-likelihood ratio tests are somewhat more sophisticated statistical distribution tests with underlying hypothesis testing.[3] These measures are widely used in CL and implemented in some corpus analysis tools, such as WordSmith Tools (Scott 1997), Wmatrix (Rayson 2009), and AntConc (Anthony 2005). One problem with these measures is that p-values tend to be very low across the board when these tests are used for comparing language corpora. The more important problem, however, is that they are designed to compare statistically independent events and handle corpora as a bag of words. These tests use the total number of words in the corpus and do not consider an uneven distribution of words within a corpus (Lijffijt et al. 2014).

Welch’s t-test, named for its creator, Bernard Lewis Welch, is an adaptation of Student’s t-test. Unlike the Student’s t-test, it does not assume an equal variance in the two populations (Welch 1947). Like the two former tests, it is also based on hypothesis testing, but in contrast to them, it takes not only the frequency of a feature into account. Sample mean, standard deviation and sample size are included in a calculation of the t-value. That is the reason why this measure can better deal with frequent words that occur only in one text or one part of a text in a given collection.

Unlike previous measures, the Wilcoxon rank sum test, also known as Mann-Whitney U-test, does not make any assumption concerning the statistical distribution of words in a corpus; in particular, it does not require the words to follow a normal distribution, as assumed by other tests such as the t-test. Corpus frequencies are usually not normally

[3] Statistical hypothesis tests are based on the computation of a p-value that expresses the probability that the observed distributions of words in a target and a comparison corpus could have arisen under the assumption that both corpora are random samples from the same underlying corpus (Oakes 1998). Put simply, such a test compares the frequency distributions of a given word in two corpora; if these distributions are very different, the probability that the two corpora are samples from the same underlying corpus is small, expressed by a small p-value, and the word is distinctive for the corpus in which it occurs more often. If, however, the distributions are very similar, then the probability that the two corpora are samples from the same underlying corpus is large, expressed by a large p-value, and the relatively small differences in the frequency distributions are most likely due to chance. The conventional threshold of statistical significance is $p = 0.05$.

distributed, making the Wilcoxon test better suited (Wilcoxon 1945; Mann and Whitney 1947; see also Oakes 1998). It is based on a comparison of a sum of rank orders of texts in two text collections. The rank orders of texts are defined according to the frequency of a target word, without considering to which of both corpora this text belongs (see Lijffijt et al. 2014). In our implementation, it sums up the frequencies per segment of documents; for this reason, we consider it to be a dispersion-based measure.

In **Computational Literary Studies**, one of the main application domains that uses measures of distinctiveness is stylometric authorship attribution. In this domain, John Burrows is famous for having introduced a distance measure he called Delta that serves to establish the degree of stylistic difference between two or several texts. (Burrows 2002). However, Burrows also defined a measure of distinctiveness, called Zeta, that was quickly taken up for concerns other than authorship (Burrows 2007). There are several variants of Zeta proposed by Craig and Kinney (2009) and by Schöch et al. (2018). Compared to measures based on statistical tests, Zeta is mathematically simple. It compares document proportions of each word in the target and comparison corpora by subtracting the two document proportion values from each other. The document proportion is the proportion of documents in the corpus in which the relevant word occurs at least once. Zeta has a bias towards medium-frequency content words. These two attributes make it attractive for other application domains in CLS, such as genre analysis (Schöch 2018) or gender analysis (Hoover 2010). Essentially, this measure quantifies degrees of dispersion of a feature in two corpora and compares them.[4] In our framework, we implemented two variants of Zeta: Burrows' Zeta (Zeta_orig, Burrows 2007) and logarithmic Zeta (Zeta_log, Schöch et al. 2018) to compare their performance.

Eta is another dispersion-based measure recently proposed by Du et al. (2021) for the comparative analysis of two corpora. Eta is based on comparing the Deviation of Proportions (DP) suggested by Gries (2008). DP expresses the degree of dispersion of a word and is obtained by establishing the difference between the relative size of each text in a corpus and the relative frequency of a target word in each text of the corpus and summing up all differences. Eta works by subtracting the DP value of a word in the target corpus from its DP value in the comparison corpus. Like Zeta, Eta therefore also compares the dispersion values of a feature, but it does so in a different way, namely, by comparing the DPs of words in two corpora.

2.2 Comparative Evaluation of Measures

The evaluation of measures of distinctiveness is a non-trivial task for the simple reason that it is not feasible to ask human annotators to provide a gold-standard annotation. Unlike a given characteristic of tokens or phrases in many annotation tasks, a given word type is distinctive for a given corpus neither in itself, nor by virtue of a limited amount of context around it. Rather, it becomes distinctive for a given corpus based on a consideration of the entire target corpus when contrasted to an entire comparison corpus. Furthermore, whether or not a word can be considered to be distinctive depends on the category that serves to distinguish the target from the comparison corpus. Commonly

[4] On dispersion, see Lyne (1985), Gries (2019) and Gries (2021b). The latter defines dispersion as "the degree to which an element – usually, a word, but it could of course be any linguistic element – is distributed evenly in a corpus" (7) and notes the unduly high correlation of most currently used dispersion measures with frequency.

used categories include genre or subgenre, authorship or author gender as well as period or geographical origin. For any meaningfully large target and comparison corpus, this is a task that is cognitively unfeasible for humans.

As a consequence, alternative methods of comparison and evaluation are required. In many cases, such an evaluation is in fact replaced by an explorative approach, based on the subjective interpretation of the word-lists resulting from two or more distinctiveness analyses, and performed by an expert who can relate the words in the word-lists to their knowledge about the two corpora that have been compared. More strictly evaluative methods (as described in more detail below) can either rely entirely on a comparison of the mathematical properties of measures (as in Kilgarriff 2001), alternatively, they can be purely statistical (as in the case of the test for uniformity of p-value distributions devised by Lijffijt et al. 2014). Finally, such an evaluation can use a downstream classification task as a benchmark (as for example in Schöch et al. 2018).

We provide some more comments on previous work in this area. Kilgarriff (2001) gives a detailed overview of statistical characteristics of some distinctiveness measures, such as log-likelihood ratio test, Wilcoxon rank sum test, t-test or TF-IDF. He suggests the χ^2 test as more suitable measure for comparative analysis, but does not provide significant empirical evidence for his claims. Paquot and Bestgen (2009) compare three measures: log-likelihood ratio test, the t-test and the Wilcoxon rank sum test. They apply these measures to find words that are distinctive of academic prose compared to fictional prose. The authors stress that the choice of a statistical measure depends on the research purpose. In the case of their analysis, the t-test showed better results, because the distribution of the words across texts in the corpus was taken into account. One of the most comprehensive evaluation studies of distinctiveness measures is provided by Lijffijt et al. (2014). The authors evaluate a wide range of measures, such as log-likelihood ratio test, χ^2 test, Wilcoxon sum rank test, t-test and others. Their evaluation strategy principally relies on a test of the uniformity of p-values designed to identify measures that are overly sensitive to slight differences in word frequencies or distributions (for details, see their paper).

Schöch et al. (2018) propose an evaluation study across two languages. They compare eight variants of Burrows Zeta by using top distinctive words as features in a classification task for assigning novels to one of two groups. According to the evaluation results, the log-transformed Zeta has the best performance; however, it remains open whether the increased performance and improved robustness come at the price of interpretability of the resulting word lists.

Egbert and Biber (2019), in turn, propose their own dispersion-based distinctiveness measure, which uses a simple measure of dispersion in combination with a log-likelihood ratio test. Its effectiveness is compared to so-called corpus-frequency methods for identifying distinctive words of online travel blogs. Their paper shows that the dispersion-based distinctiveness measure is better suited compared to the other measures. Their paper, however, is lacking a systematic comparison of the new measure to other established measures of distinctiveness and does not really provide a significant empirical evaluation of their method.

Du et al. (2021), finally, provide a comparison of two dispersion-based measures, namely Zeta and Eta, for the task of extracting words that are distinctive of several subgenres of French novels. The authors come to the conclusion that both measures are able to identify meaningful distinctive words for a target corpus compared to another corpus but do not consider a usefully broad range of measures.

Concerning an evaluation across languages, to the best of our knowledge, evaluations of measures of distinctiveness that use corpora in more than one language are virtually non-existent. The only example that comes to our mind is Schöch et al. (2018) who used a Spanish and a French corpus for evaluation but only provide detailed information on the results for French. Unless we have missed relevant publications, our contribution is the first study that includes an evaluation of measures of distinctiveness on corpora in multiple languages.

3. Corpora

For our analysis we used nine text collections. The first two corpora consist of contemporary popular novels in French published between 1980 and 1999 (160 novels published in the 1980s and 160 novels published in the 1990s). To enable the comparison and classification of texts, we designed these custom-built corpora in a way that they contain the same number of novels for each of four subgroups: highbrow novels on the one hand, and lowbrow novels of three subgenres (sentimental novels, crime fiction and science fiction) on the other. The texts in these corpora are, for obvious reasons, still protected by copyright. As a consequence, we cannot make these corpora freely available as full texts. We have published them, however, in the form of a so-called "derived text format" (see Schöch et al. 2020; Organisciak and Downie 2021) suitable for use with our Python library and devoid of any copyright protection.[5]

Another group of text corpora that we used for our analysis consists of seven collections of novels in seven different European languages taken from the *European Literary Text Collection* (ELTeC) produced in the COST Action *Distant Reading for European Literary History* (see Burnard et al. 2021; Schöch et al. 2021).[6] We reuse the English, French, Czech, German, Hungarian, Portuguese and Romanian corpora. From each of these corpora, we selected a subset of 40 novels: 20 novels from the period from 1840 to 1860 and 20 novels from the period from 1900 to 1920.

[5] See `https://github.com/Zeta-and-Company/derived-formats`; DOI: 10.5281/zenodo.7111522.
[6] Texts and metadata for these collections are available on Github: `https://github.com/COST-ELTeC`; DOI: 10.5281/zenodo.3462435. On the COST Action more generally, see also: `https://www.distant-reading.net/`.

corpus		document length		number of	
name	size (million words)	standard deviation	mean	types	authors
fra_80s	8.83	27,161	55,225	119,775	120
fra_90s	8.48	26,976	53,010	111,501	124
ELTec_cze	1.98	24,734	49,642	163,900	33
ELTec_deu	4.62	101,915	115,531	158,726	30
ELTec_eng	4.66	75,672	116,477	53,285	35
ELTec_fra	3.31	86,926	82,802	65,799	37
ELTec_hun	2.44	40,513	61,055	258,026	36
ELTec_por	2.33	38,787	58,325	95,572	34
ELTec_rom	2.41	36,493	60,395	156,103	37

Table 2: Overview of the corpora used in our experiments.

4. Methods

To obtain a better understanding of the performance of different measures of distinctiveness, we evaluate how well the words selected by these measures are helpful for distinguishing texts into predefined groups. As mentioned above, we focus on subgenre (and, to a lesser degree, on time period) as the distinguishing category of these text groups here because these are both highly relevant categories in Literary Studies. This means that among the approaches for comparative evaluation outlined above, we have adopted the downstream classification task for the present study. The main reasons for this choice are that the rationale and the interpretation of this evaluation test is straightforward and that it can be implemented in a transparent and reproducible manner. In addition, we assume that it will give us an idea of how suitable the different measures are for identifying the words that are in fact distinctive of these groups.

In order to identify distinctive words, we first define a target corpus and a comparison corpus and run the analysis using nine different measures, including two variants of the Zeta measure. Concerning the first two corpora, which consist of contemporary French novels, we are interested in distinctive words for each of the four subgenres. Concerning the second, multilingual set of corpora, we make a separate comparison for each language based on two periods: earlier vs. later texts.

For the distinctiveness analysis of the contemporary French novels, we took novels from each subgenre as the target corpus and the novels from the remaining three subgenres as the comparison corpus. This means that we ran the distinctiveness analysis four times and obtained four lists of distinctive words for each subgenre and another four lists of distinctive words for each comparison corpus (words that are not 'preferred' by the target corpus). For the classification of these novels, which is a four-class classification scenario, we took the N most distinctive words from each of the above-mentioned eight lists to classify the documents. Therefore, $N \times 8$ features are actually used for the classification tasks.

For the multilingual set of corpora, the situation is simpler, because there are only two classes. We can get two lists of words, which are the distinctive words for each class by running the distinctiveness analysis only once, which takes one class (novels from 1840 to 1860) as the target corpus and the other class (novels from 1900 to 1920) as the comparison corpus. Here, we also took the N most distinctive words from each of these two lists to classify the documents. Therefore, $N \times 2$ features are actually used for the classification tasks.

To observe the impact of N on the classification performance, we classify corpora using different settings of $N \in \{10, 20, 30, 40, 50, 100, 200, 300, 400, 500, 1000, 2000, 3000, 4000, 5000\}$. Based on the absolute frequency of these features, we perform a classification task. As explained above, as classification units we do not use the entire novels, but segments of 5000 words. As the classification accuracy measure, we use the F1-score (F1-macro mean). The performance is evaluated in a 10 fold cross-validation setting.

In order to create a baseline for the classification tasks, we randomly sample $N \times 8$ words from each of the two French novel collections and $N \times 2$ words from each corpus of the multilingual collection and perform the segment classification based on the absolute frequency of these words. This process has been repeated 1000 times and the mean F1-score is defined as the baseline.

5. Results

5.1 Classification of French Popular Novel Collections (1980s and 1990s)

This section describes the classification of French novel segments into four predefined classes: highbrow, sentimental, crime and scifi. Before running the tests on the corpora of different languages, we want to check the variance of results within one language. Only by excluding one confounding variable (language) from the test, we can conclude that the differences in the performance of measures of the ELTeC-corpora are caused by the differences among different languages. That's why we built two corpora of French novels for our analysis: novels from the 1980s and from the 1990s.

First we applied bag-of-words based classification on both parts of the French novel corpus, testing four classifiers: Linear Support Vector Classification, multinomial Naive Bayes (NB), Logistic Regression and Decision Tree Classifier.[7]

Figure 1 shows the classification results of the 1980s-corpus. The Decision Tree Classifier has a clearly lower performance than the other three classifiers. The other three classifiers produce better results with similar trends of F1-scores across different measures. Therefore, in our further experiments we focus on results based on one classifier, namely the Multinominal NB.[8] The classification results of the 1990s-corpus, for this preliminary test, are very similar to the results presented in Figure 1 and thus are not shown here.

Figure 2 shows the F1-macro score distribution from 10 fold cross-validation for classification of the French novel segments of the 1980s-dataset. The setting of N varies from 10 to 5000. The baseline is visualized as a green line in the plot. It corresponds to the average of the classification results based on $N \times 8$ random words, resampled 1000 times.

The classification based on the N most distinctive features leads almost always to better

[7] LinearSVC, MulinomialNB, LogisticRegression and DecisionTreeClassifier from the Python package *scikit-learn*; see: https://scikit-learn.org/.

[8] According to https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html, Naive Bayes methods are suggested for classification of text data.

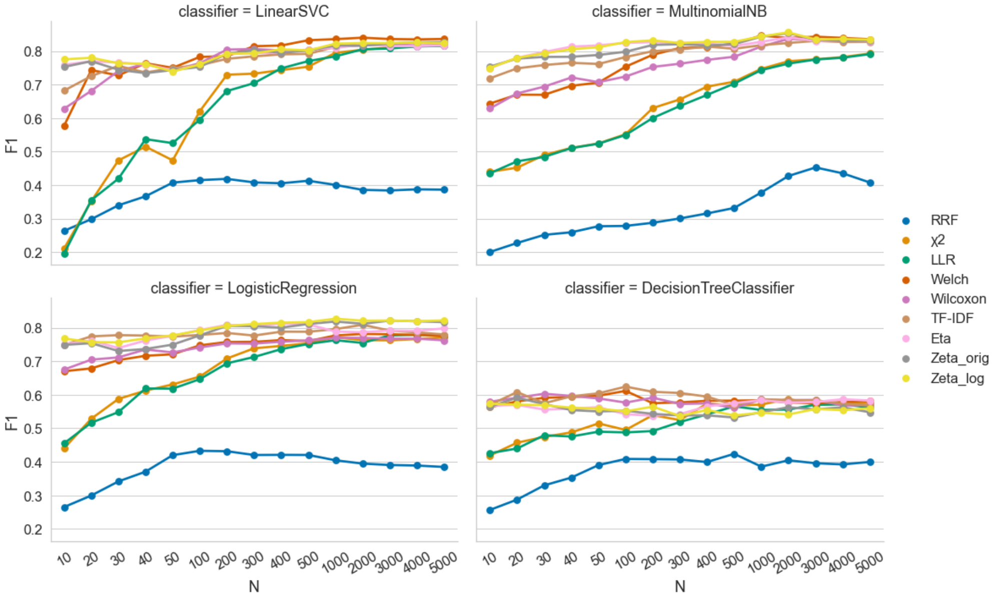

Figure 1: Classification performance on the French corpus (1980s) with four classifiers, depending on the distinctiveness measure and the setting of N.

classification results, compared to the baseline. The smaller the number of features, the bigger is the difference between the F1-scores of the baseline and those of the classifiers. The baseline approaches the performance of the classifier that uses distinctive words as the number of features increases. This can be explained, firstly, by the continuously increasing baseline performance. Secondly, we observe that with a high number of features, almost all measures have similarly high F1-scores. Thirdly, we assume, all lists of distinctive words become more and more similar to each other and have considerable overlap with the vocabulary of the segments at some point. Interestingly, however, as we can see in the Figures 1 and 2, some measures (among them both Zeta variants, Eta, Wilcoxon and Welch) almost constantly perform with high F1-scores that are clearly above the baseline, even when the classification is performed with only $N = 10$ features.

Another observation based on Figure 2 is that the differences in the variations of the F1-score distributions decrease with the increase of N. The measures also show different degrees of variation of results depending on the corpora.[9] In order to identify which distinctiveness measures produce features that lead to results that are significantly better and more robust, we applied a two-tailed t-test on every pair of the F1-score distributions. The results for the 1980s text collection are shown in Figure 3.

In Figure 3, each boxplot represents the distribution of 36 p-values (all pairwise combinations of nine measures) at the setting of the corresponding N. We can observe that with increasing N, the number of p-values smaller than 0.05 (significance threshold) decreases.[10] This means that the more features are used, the less statistically significant differences exist between classification results. This observation proves our previous conclusion, that a high number of features automatically leads to high accuracy and (certainly, according to the p-values, from $N = 3000$) it is not important, in such a

[9] Classification of the 1980s-collection leads to lower variations of the F1-scores compared to the classification of the 1990s-collection.
[10] When $N = 10 - 100$, more than 50% of the p-values are below the threshold of 0.05 and when $N = 300$ or higher, most of the p-values are above the threshold of 0.05.

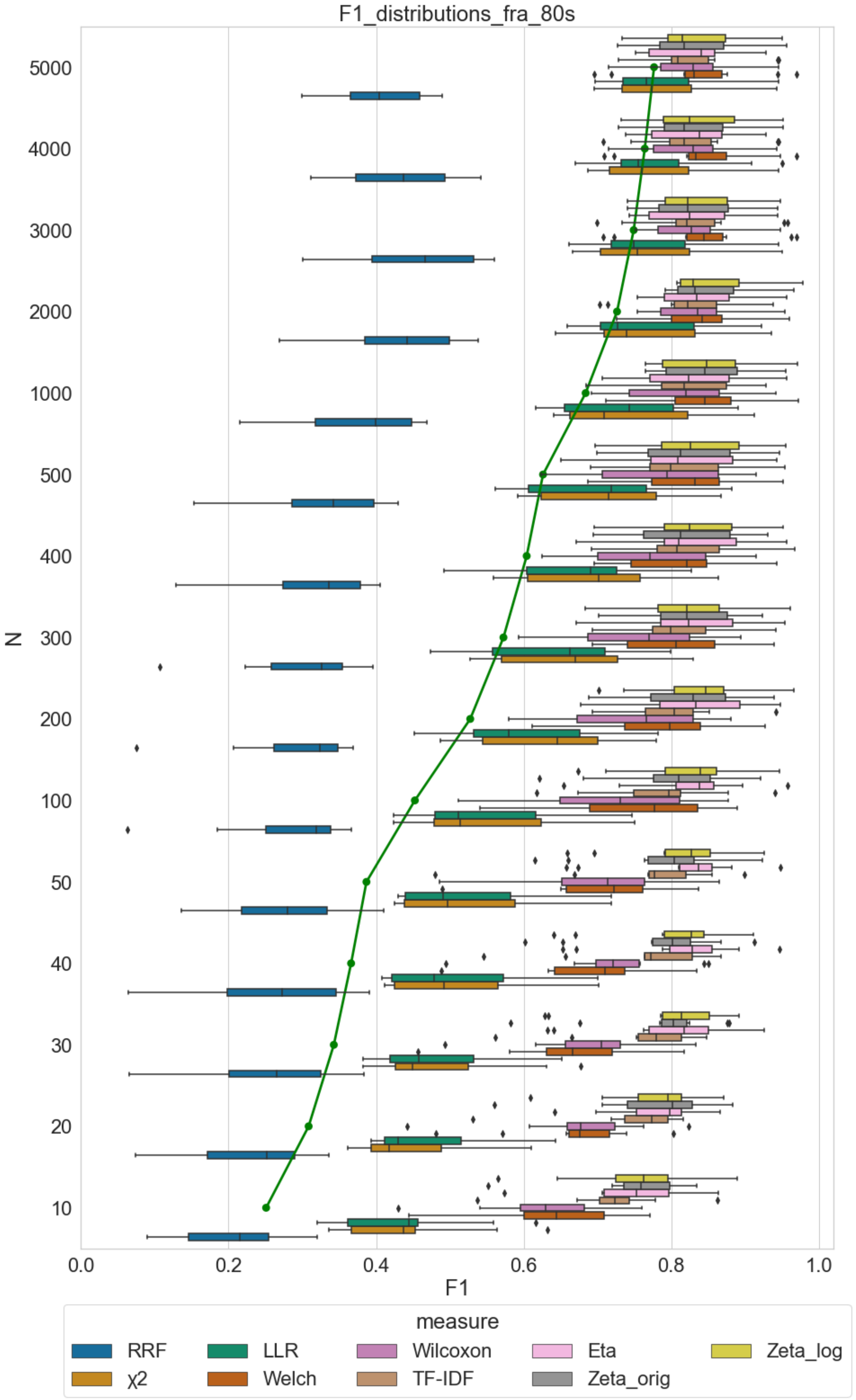

Figure 2: F1-macro score distribution from the 10 fold cross-validation obtained by the genre classification of the French 1980s-corpus with Multinominal NB. The green line is the baseline F1-score.

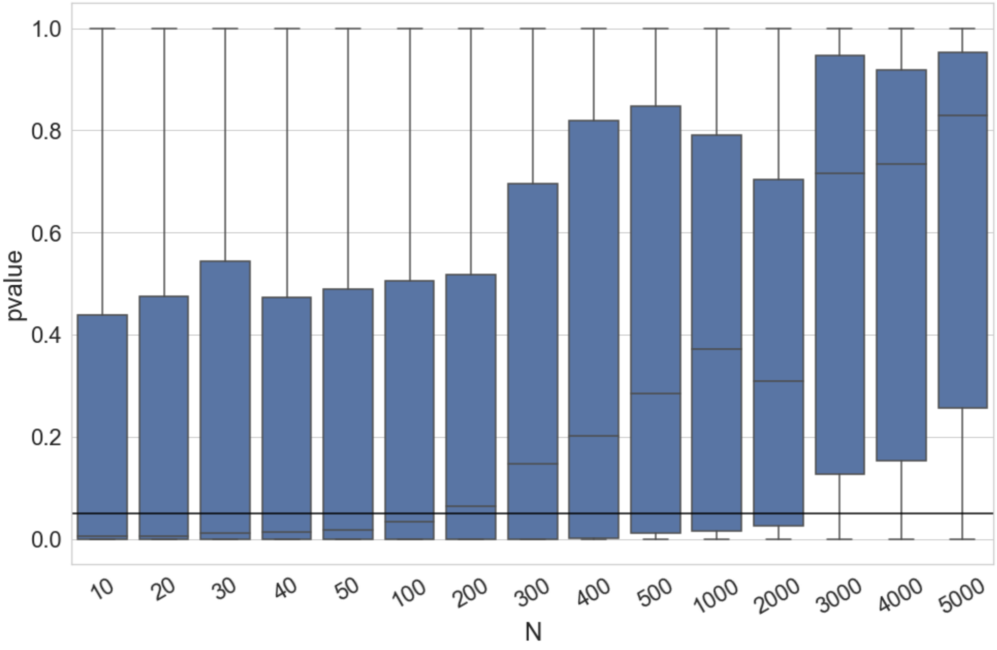

Figure 3: T-test performed on every pair of the F1-score distributions of measures. F1-score were obtained from the classification of the 1980s-corpus. The black line is the significance threshold.

scenario, which measure is used.

The more interesting observation, however, is that we have clear differences in F1-scores of the measures when a small number of features is used (e.g. $N = 10, 20, 30$).[11] To investigate this phenomenon in more detail, we visualized heatmaps with p-values obtained from a t-test on pairs of the F1-score distributions of measures for the classification with $N = 10$ features only (Figure 4).

First of all, we can observe in Figure 4(a) that for the classification with $N = 10$ features, the F1-scores of RRF, χ^2 and LLR are very low, Wilcoxon and Welch have average performance, while both Zeta variants, Eta and TF-IDF have the highest scores.[12]

We can also observe, in Figure 4(b), that RRF is an outlier and has significantly different F1-scores compared to all other measure. χ^2 and LLR have almost perfect correlation with each other and significantly differ from all other measures as well as from RRF. We can make the same observation concerning the Wilxocon and Welch measures: they have a strong correlation with each other and significantly different results to other measures with exception of TF-IDF. As for the other measures, we observer a high correlation in F1-scores between TF-IDF, Eta and both Zetas. Combining this information with F1-score distributions at $N = 10$ (Figure 4a) lets us affirm that all frequency-based measures (RRF, LLR and χ^2) perform significantly worse compared to the other measures, when we set $N = 10$ for our classification task. Concerning both Zetas, Eta and TF-IDF we can conclude that they have significantly better results compared to other measures.

[11] This observation on the 1980s-dataset can also be seen in the results from tests on the 1990s-dataset.
[12] RRF median = 0.22, χ^2 = 0.44, LLR = 0.45, Wilcoxon = 0.63, Welch = 0.65, TF-IDF = 0.73, Eta = 0.76, Zeta_orig = 0.77, Zeta_log = 0.77.

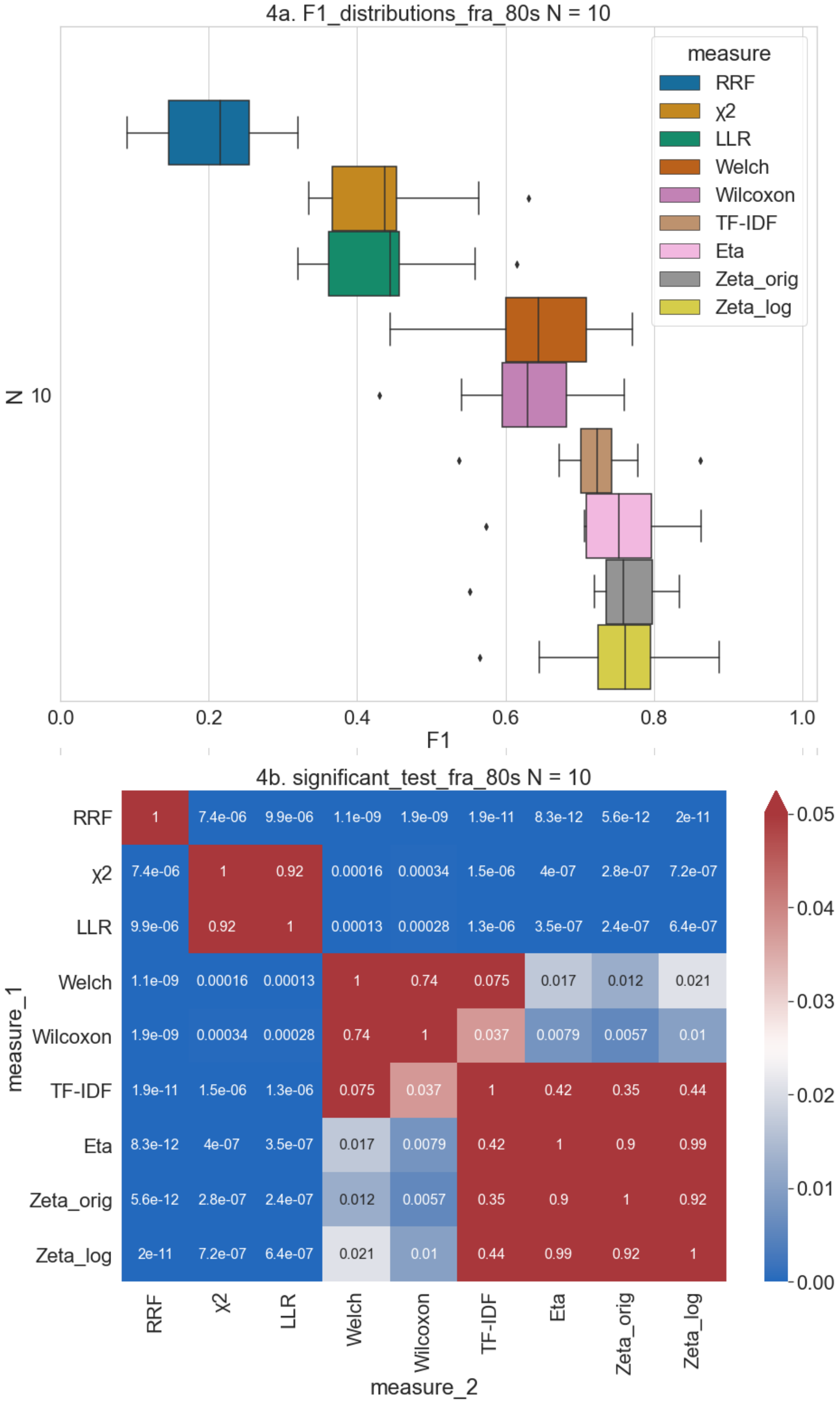

Figure 4: (a) F1-score distributions for classification with $N = 10$. (b) p-values obtained from t-tests on pairs of the F1-score distributions of measures. F1-scores obtained from the classification of the 1980s-corpus with $N = 10$. Significance threshold is 0.05. Note that all values above 0.05 are shown in red.

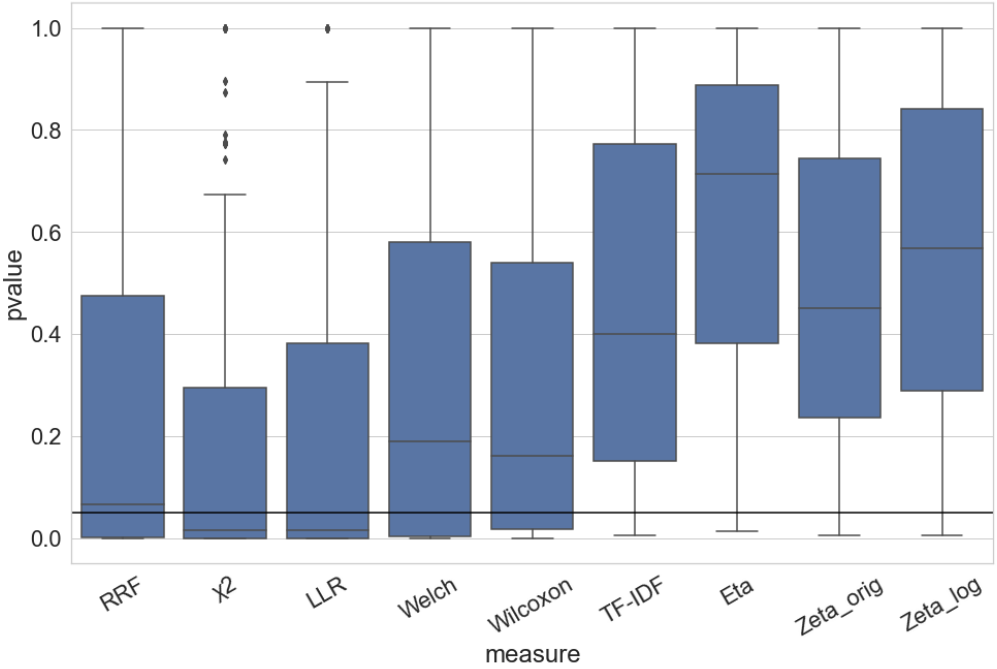

Figure 5: Significance test on F1-score distributions for each measure. F1-scores obtained from the classification of the 1980s-corpus. Black line is significance threshold.

Wilcoxon and Welch have average performance and similar scores, a fact that explains their relatively high correlation.[13]

This observation applies for classifications with greater N as well. We can also note, however, that results in these cases are not stable and have a high variation of F1-score distributions depending on the coice of N and the corpus. In order to ascertain whether these variations in results are significant and which measures perform with robustly high F1-scores, we also analyzed the classification results within each measure through significance tests on F1-score distributions (Figure 5). The results of the significance tests with p-values below the threshold of 0.05 would mean that the differences in the F1-scores are significant and did not occur by chance. On the other hand, p-values above the threshold mean that there are only slight, insignificant differences in F1-scores. If the F1-scores only show little variation, this also means that the performance of the measure is stable and robust.

Figure 5 shows that almost all p-values obtained from the F1-scores of both Zeta variants, Eta and TF-IDF are greater than the significance threshold of 0.05. The Wilcoxon and Welch have around 25% of p-values below 0.05. This means that the classification results based on features extracted by these measures are stable and robust, independently of the choice of N. Concerning LLR and χ^2, there are over 50% of p-values below the significance threshold, RRF has around 50% of p-values below 0.05.[14]

Summarizing the information from the classification of both corpora, we can argue that Zeta_log, Zeta_orig, Eta and TF-IDF have the highest and the most robust performance

[13] We observe a slightly different tendency for the classification of the 1990s-dataset: Both Zetas, Eta, TF-IDF, Welch and Wilcoxon do not have significant differences in F1-scores for $N = 10$.
[14] The results of the classification of the 1990s-dataset show the same tendency.

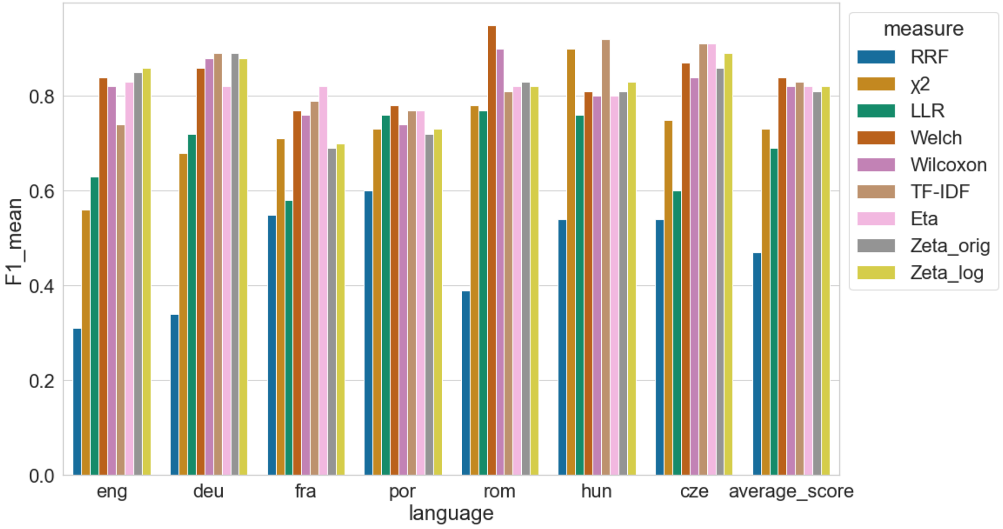

Figure 6: Mean F1-score of classification across 7 ELTeC corpora ($N = 10$).

when using the smallest number of features ($N = 10$).[15] These results mean that 10 words identified as distinctive by these measures are sufficient to correctly distinguish over 70% of texts into four groups.

It is important to note that this group of the most successful measures have something in common: they are all dispersion-based (TF-IDF with some restrictions).[16] It appears fair to conclude that in our case, dispersion-based measures can best identify the words that are the most distinctive for a certain genre. The frequency-based measures show a significantly lower and less stable performance. Wilcoxon and Welch show average results.[17]

5.2 Experiments on Seven ELTeC Text Collections

The above-mentioned conclusion regarding the superior performance of dispersion-based measures when compared to frequency-based measures is based on the specific use-case of our 20th-century French novel corpus. In order to verify whether this claim is also true when corpora in other languages are used, we performed the same tests on several subsets derived from ELTeC (as described above, Section 3), namely from the English, French, Czech, German, Hungarian, Portuguese and Romanian collections.

The classification task that we use differs from the previous one. We are not interested in classifying the texts by subgenre, but by their period of first publication (1840-1860 vs. 1900-1920). The main reason for this is practical: the corpora included in ELTeC do not have consistent metadata regarding the subgenre of the novels included, due to the large variability of definitions and practices in the various literary traditions that are covered by ELTeC. However, all collections cover a very similar temporal scope so that it is possible to use this as a shared criterion to define two groups for comparison.

[15] Zeta_log has the highest mean F1-score (1980s: 0.75, 1990s: 0.72), followed closely by Eta (1980s: 0.75, 1990s: 0.72), and then by Zeta_orig (1980s: 0.75, 1990s: 0.70), TF-IDF (1980s: 0.72, 1990s: 0.71).
[16] Dispersion describes the even/uneven spread of words across a corpus or across each particular text in a corpus. We cannot claim, however, that the measures we have used rely exclusively on dispersion; rather, they are also influenced by frequency; see Gries (2021b).
[17] For information about the types of measures, see Table 1.

We consider the performance across corpora and measures for $N = 10$, based on the mean F1-score of the classification task (Figure 6). We can observe that among the tests based on seven corpora, five of them could achieve a result of 0.8 or higher. In particular, the dispersion-based and the distribution-based measures can guarantee good or even best results in almost every classification task. The only exception is the classification of the Portuguese corpus. The classification results based on other measures are very similar, except for RRF. Both Zeta variants and Eta are among the best classification results for the English, German, Hungarian and Czech corpora, while Welch and TF-IDF yielded particularly good results when classifying the Romanian corpus.

With regard to the frequency-based measures, we can observe that χ^2 has very good results for the Hungarian corpus, but not for the English or German corpora. LLR has relatively high scores for the Portuguese and Hungarian corpora. But in most cases, it is still not as good as dispersion-based measures such as Zeta_log. Compared to all other measures, the ten most distinctive words defined by the RRF lead to the worst results in all classification tasks.

Considering further analyses, we visualized[18] the difference between the F1-score distributions for each measure with varying N. In a similar way to the results from the French novel sets, the differences decrease with increasing N. However, unlike the results from the French novel sets in Figure 3, some corpora have more than 75% of the p-values greater than 0.05 when N is greater than 100 (e.g. Czech and German corpora). Some do not have the same results until N is greater than 500 (e.g. English corpus). This indicates that, although the results show some variations between the different corpora, the overall trend is the same. The larger the value of N, the less important it is which measure is used to select the features (distinctive words) for classification.

If we consider the stability of the measures across evaluation with different numbers of features, we can conclude that the results for several measures (RRF, Welch, Wilcoxon, ETA, Zeta_orig and Zeta_log) are stable: for almost all data sets, the number of significantly different results is less than 25%. This indicates that the setting of N has little effect on the results of the classification. Increasing N does not significantly improve the classification results. This suggests that these measures (except RRF, which does not deliver good results in any classification task, regardless of how N is set) can work well to find those most distinctive features. As for frequency-based measures, we have a contrary observation: In most cases, the results of the classification are significantly different with different settings of N.

Summarizing the results described above, we can conclude that dispersion-based and distribution-based measures have been shown again to yield higher performance in identifying distinctive words and to be more stable and robust than other measures. In contrast, the average performance of frequency-based measures is still considerably lower than that of the other measures.

[18] The data is available in our GitHub repository: https://github.com/Zeta-and-Company/JCLS2022/tree/main/Figures.

6. Conclusion and Future Work

To conclude, we have been able to show that a Naive Bayes classifier performs significantly better in two different classification tasks when it uses a small number of features selected using a dispersion- or distribution-based measure, compared to when it uses a small number of features selected using a measure based on frequency. This result was quite robust across all nine different corpora in seven different languages. In addition, we were able to observe it both for the four-class subgenre classification tasks and the two-class time period classification task. In this sense, our findings support an emerging trend (see e.g. Egbert and Biber 2019; Gries 2021a) to consider dispersion to be an important property of words in addition to frequency.

However, this result also comes with a number of provisos: We have observed this result only for small values of N: in fact, the advantage of the dispersion-based measures decreases as the number of features increases. In addition, we have observed this result for classification tasks in which a small segment of just 5000 words needed to be classified. We suspect, but have not verified this hypothesis for the moment, that this advantage may be reduced for larger segments. For the moment, finally, we have not yet systematically verified whether the same results can be obtained for classifiers other than the one used in our experiments.

The fact that these results can only be observed for small values of N, disappearing for larger values of N, is noteworthy. In our opinion, this does not mean that the best solution is to use larger values of N and stop worrying about measures of distinctiveness altogether. The main reason, we believe, why using smaller values of N is useful, in addition to the general principle of Occam's razor, is related to interpretability: Regardless of the interpretability of the individual words they are composed of, the interpretability of word lists decreases with increasing values of N, simply because it becomes increasingly challenging to intellectually process and interpret word lists growing much beyond 100 items.

Despite these results, there are of course a number of issues that we consider unsolved so far and that we would like to address in future work. The first issue was already mentioned above and concerns the length of the segments used in the classification task. As a next step, we would like to add *segment length* as a parameter to our evaluation pipeline in order to test the hypothesis that the advantage of dispersion-based measures disappears for segments substantially longer than 5000 words.

The second issue concerns the number and range of measures of distinctiveness implemented in our Python package so far. With nine different measures, we already provide a substantial number of measures. However, we plan to add several more measures to this list, notably Kullback-Leibler Divergence (a distribution-based measure, see: Kullback and Leibler 1951), the measure combining dispersion and log-likelihood ratio used by Egbert and Biber (2019), the 'inter-arrival time' measure proposed by Lijffijt et al. (2011), a measure yet to be defined that would be based on the 'pure' dispersion measure DP_{nofreq} recently proposed by Gries (2021b) as well as the LRC measure proposed by Evert (2022).

Thirdly, it should be considered that almost all previous studies in the area of distinctiveness, our own included, do not allow any conclusions as to whether the words defined by a given measure as statistically distinctive are also perceived by humans as distinctive. Such an empirical evaluation is out of scope for our paper, but would certainly add a different kind of legitimacy to a measure of distinctiveness. In addition, words that prove to be statistically distinctive in a classification task are, strictly speaking, only shown to have a certain discriminatory power in the setting defined by the two groups of texts. Distinctiveness, however, can be understood in more ways than just discriminatory power; notably, distinctiveness can also be understood in terms of salience or aboutness.[19]

Finally, we would of course like to expand our research regarding the elephant in the room, so to speak: not just evaluating statistically which measures perform more or less well in particular settings, but also explaining why they behave in this way. We believe that the distinction between measures based on frequency, distribution and dispersion is a good starting point for such an investigation, but pushing this further also requires to include measures that really measure only dispersion and not a mix of dispersion and frequency, as recently demonstrated by Gries (2021b). Measures of distinctiveness have clearly not yielded all their secrets to us yet.

7. Data Availability

Data can be found here: https://github.com/Zeta-and-Company/JCLS2022 (DOI: https://doi.org/10.5281/zenodo.6517748).

8. Software Availability

Software can be found here: https://github.com/Zeta-and-Company/JCLS2022 (DOI: https://doi.org/10.5281/zenodo.6517748).

9. Acknowledgements

The research reported here was conducted in the framework of the project ‘Zeta and Company’ funded by the German Research Foundation (DFG, project number 424211690). This project is part of the Priority Programme SPP 2207 ”Computational Literary Studies”. We thank our reviewers for their valuable feedback to an earlier version of this paper.

10. Author Contributions

Keli Du: Methodology, Investigation, Visualization, Software, Writing - review & editing

Julia Dudar: Formal Analysis, Data curation, Writing - original draft, Writing - review & editing

[19] For a theoretical take on both issues raised here, see Schröter et al. (2021).

Christof Schöch: Conceptualization, Data curation, Funding acquisition and Supervision, Writing - original draft, Writing - review & editing

References

Anthony, Laurence (2005). "AntConc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom". In: *Proceedings, International Professional Communication Conference, 2005 (IPCC 2005)*, 729–737. 10.1109/IPCC.2005.1494244.

Burnard, Lou, Christof Schöch, and Carolin Odebrecht (2021). "In search of comity: TEI for distant reading". In: *Journal of the Text Encoding Initiative* 14. 10.4000/jtei.3500.

Burrows, John (2002). "'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship". In: *Literary and Linguistic Computing* 17 (3), 267–287. 10.1093/llc/17.3.267.

— (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.

Craig, Hugh and Arthur F. Kinney, eds. (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge University Press.

Damerau, Fred J. (1993). "Generating and evaluating domain-oriented multi-word terms from texts". In: *Information Processing & Management* 29 (4), 433–447.

Diez, David, Mine Cetinkaya-Rundel, and Christopher D. Barr (2019). *OpenIntro Statistics*. 4th ed. OpenIntro. https://www.openintro.org/book/os/ (visited on 10/24/2022).

Du, Keli, Julia Dudar, Cora Rok, and Christof Schöch (2021). "Zeta & Eta: An Exploration and Evaluation of Two Dispersion-based Measures of Distinctiveness". In: *Computational Humanities Research 2021* (*CEUR Workshop Proceedings*). Ed. by Maud Ehrmann, Folgert Karsdorp, Melvin Wevers, Tara Lee Andrews, Manuel Burghardt, Mike Kestemont, Enrique Manjavacas, Michael Piotrowski, and Joris van Zundert. CEUR. http://ceur-ws.org/Vol-2989/ (visited on 10/24/2022).

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 10/24/2022).

Egbert, Jesse and Doug Biber (2019). "Incorporating text dispersion into keyword analyses". In: *Corpora* 14 (1), 77–104. 10.3366/cor.2019.0162.

Evert, Stephanie (2022). "Measuring keyness". In: Publisher: OSF. 10.17605/OSF.IO/CY6MW.

Gries, Stefan Th. (2008). "Dispersions and adjusted frequencies in corpora". In: *International Journal of Corpus Linguistics* 13 (4), 403–437. 10.1075/ijcl.13.4.02gri.

— (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.

— (2019). "Analyzing dispersion". In: *A Practical Handbook of Corpus Linguistics*. Ed. by Magali Paquot and Stefan Th. Gries. Springer, 99–118.

— (2021a). "A new approach to (key) keywords analysis: Using frequency, and now also dispersion". In: *Research in Corpus Linguistics* 9, 1–33. 10.32714/ricl.09.02.02.

— (2021b). "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.

Hempfer, Klaus W. (2014). “Some Aspects of a Theory of Genre”. In: *Linguistics and Literary Studies / Linguistik und Literaturwissenschaft*. Ed. by Monika Fludernik and Daniel Jacob. De Gruyter. 10.1515/9783110347500.405.

Hoover, David L. (2010). “Teasing out Authorship and Style with t-tests and Zeta”. In: *Digital Humanities Conference (DH2010)*. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-658.html (visited on 10/24/2022).

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

Klimek, Sonja and Ralph Müller (2015). “Vergleich als Methode? Zur Empirisierung eines philologischen Verfahrens im Zeitalter der Digital Humanities”. In: *Journal of Literary Theory* 9 (1). 10.1515/jlt-2015-0004.

Kullback, Solomon and Richard A. Leibler (1951). “On information and sufficiency”. In: *The annals of mathematical statistics* 22 (1), 79–86.

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.

Lijffijt, Jefrey, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila (2011). “Analyzing Word Frequencies in Large Text Corpora Using Inter-arrival Times and Bootstrapping”. In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis. Springer, 341–357.

Luhn, Hans Peter (1957). “A statistical approach to mechanized encoding and searching of literary information”. In: *IBM Journal of research and development* 1 (4), 309–317.

Lyne, Anthony A. (1985). “Dispersion”. In: *The Vocabulary of French Business Correspondence: Word Frequencies, Collocations and Problems of Lexicometric Method*. Slatkine, 101–124.

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.

Oakes, Michael P. (1998). *Statistics for Corpus Linguistics*. Edinburgh University Press.

Organisciak, Peter and J. Stephen Downie (2021). “Research access to in-copyright texts in the humanities”. In: *Information and Knowledge Organisation in Digital Humanities*. Routledge, 157–177. 10.4324/9781003131816-8.

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 and Rodopi. 10.1163/9789042029101_014.

Rayson, Paul (2009). *Wmatrix: a web-based corpus processing environment*. Computing Department, Lancaster University.

Salton, Gerard and Christopher Buckley (1988). “Term-weighting approaches in automatic text retrieval”. In: *Information Processing & Management* 24 (5), 513–523. 10.1016/0306-4573(88)90021-0.

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 (2022). *Computational Genre Analysis*. https://dragonfly.hypotheses.org/1219 (visited on 11/23/2022).

Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, and Jörg Röpke (2020). "Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen". In: *Zeitschrift für digitale Geisteswissenschaften*. 10.17175/2020_006.

Schöch, Christof, Roxana Patras, Tomaž Erjavec, and Diana Santos (2021). "Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives". In: *Modern Languages Open* 1, 1–19. 10.3828/mlo.v0i0.364.

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*. ADHO. https://dh2018.adho.org/burrows-zeta-exploring-and-evaluating-variants-and-parameters/ (visited on 10/24/2022).

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 (JLT)* 9 (1–2), 81–108. 10.1515/jlt-2021-2011.

Scott, Mike (1997). "PC Analysis of Key Words and Key Key Words". In: *System* 25 (2), 233–245.

Spärck Jones, Karen (1972). "A statistical interpretation of term specificity and its application in retrieval." In: *Journal of Documentation* 28, 11–21.

Tufte, Edward R. (1990). *Envisioning information*. Graphics Press.

Welch, B. L. (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), 80–83. 10.2307/3001968.