

## Expertise vs. statistics. A qualitative evaluation of three keyness measures (logarithmic Zeta, Welch's t-test, and Log-likelihood ratio test) applied to subgenres of the French novel

Julia Röttgermann <roettger\_at\_uni-trier\_dot\_de>, University Trier, Germany  <https://orcid.org/0000-0002-1918-8117>

Keli Du <duk\_at\_uni-trier\_dot\_de>, University of Trier, Germany  <https://orcid.org/0000-0001-7800-0682>

Julia Havrylash <julia\_dot\_havrylash\_at\_uni-trier\_dot\_de>, University of Trier, Germany  <https://orcid.org/0000-0001-5545-9562>

Christof Schöch <schoech\_at\_uni-trier\_dot\_de>, University of Trier, Germany  <https://orcid.org/0000-0002-4557-2753>

DOI: <https://doi.org/10.63744/djrqs3es4czs>

### Abstract

This paper continues an ongoing investigation of measures of distinctiveness (also known as keyness measures), this time employing a qualitative, comparative evaluation of three different measures: logarithmic Zeta, Welch's t-test, and Log-likelihood ratio test. Our domain of application is the contemporary French novel, more specifically four types of French novels from the period 1970-1999, namely: sentimental novel, crime novel, science fiction, and "littérature blanche" (literary fiction).

Our evaluation proceeds in the following steps: First, we establish important abstract characteristics of specific literary subgenres based on a synthesis of close readings of scholarly literature on these subgenres, resulting in qualitative, expert-based "subgenre profiles." Second, we use a purely statistical approach, namely three different measures of distinctiveness, to identify words that are expected to be statistically typical or characteristic of groups of texts such as subgenres, when compared to other texts. Finally, we compare expertise and statistics, that is, attempt to establish, for each of the four subgenres, a mapping between individual words found to be statistically distinctive of this subgenre and specific aspects contained in the relevant subgenre profile and count the matches.

It turns out that each measure yields a different list of most distinctive words that therefore, relates differently to the subgenre profiles. The analysis of these varying degrees of overlap contributes to a better understanding of the characteristics of and differences between the three measures, while also serving as an example of a qualitative evaluation of a statistical measure.

## Introduction

The identification of distinctive words for a certain group of texts has become a common task in corpus linguistics and computational literary studies. Therefore, measures of distinctiveness lie at the core of many widely used digital humanities tools, from keyword analysis in corpus software (e.g., TXM [Heiden 2010], AntConc [Anthony 2024], WordSmith Tools [Scott 2024]) to stylometric frameworks (e.g., Stylo for R [Eder et al, 2016]). These tools typically rely on established statistical procedures to calculate what counts as "key" or "distinctive," often defaulting to a certain frequency- or distribution-based measure, e.g. the log-likelihood ratio test. However, such defaults are rarely questioned, despite their significant impact on the results of literary and linguistic analysis. The present study addresses this issue by comparing different statistical measures for extracting distinctive words from subgenres of literary corpora. It has been conducted in the framework of a project in which several statistical measures to identify distinctive words – based on the frequency, the distribution and/or the dispersion of words in corpora – are being investigated.<sup>[1]</sup> For this

purpose, we created a corpus consisting of several different sub-genres of the French novel of the 1970s, 1980s, and 1990s – namely science fiction, sentimental novel, detective novel, and “littérature blanche” (literary fiction).<sup>[2]</sup> The corpus was created with the aim of comparing distinctive words in different genres, as well as comparing different statistical measures for extracting distinctive words. In a previous study, we have compared nine different keyness measures quantitatively [Du et al. 2022b]. For the present study, we decided to select one very well-performing measure from each of three groups of measures: logarithmic Zeta as a dispersion-based measure [Burrows 2007] [Schöch 2018], Welch’s t-test as a distribution-based measure [Welch 1947], and log likelihood ratio test as a frequency-based measure [Dunning 1993].

## Previous work

The statistical evaluation of word frequencies and distributions has a long history in linguistics, dating back (at least) to Yule’s analysis of noun frequencies in his 1944 study “De Imitatione Christi” [Yule 1944, p.3]. Yule’s aim was to use these statistical evaluations to determine authorship, an application of word frequency studies often called “stylometry” that remains popular and in active development to this day [Burrows 2007] [Craig and Kinney 2009] [Kestemont 2018] [Savoy 2020] [Rebora 2023] [Nagy 2024]. The goal in our research project, on the other hand, is to extract distinctive words per subgenre and to gain a deeper understanding of the performance and particularities of different measures of distinctiveness.

In the realm of automatic genre detection, Brett Kessler, Geoffrey Nunberg, and Hinrich Schütze (1997) have made an attempt to correlate genre with various linguistic surface cues. These cues include a range of structural, lexical, character-level, and derivative features. Structural cues involve passives, nominalizations, and topicalized sentences, while lexical cues include terms of address or Latinate affixes, and character-level cues encompass punctuation cues and other separators. By using these linguistic surface cues, their approach aims to automatically identify the genre [Kessler et al. 1997, p. 33–38] and brow (high brow, uppermiddle brow, middle brow, popular) of a given text using logistic regression and neural networks. Theoretically, distinctive words should also be useful lexical cues.

Expanding on the findings of Kessler, Nunberg and Schütze (1997), Joseph Worsham and Jugal Kalita (2018) have worked on the problem of classification-based genre identification but update it with more recent methods like deep learning and text classification models. They emphasize the unique challenge of computational literary studies: “Rather than communicating succinct and directed information, literature is artistic and conveys complicated themes over the course of very long narratives” [Worsham and Kalita 2018, p. 163]. Using data from Project Gutenberg with six different subject headings – Science fiction, Adventure stories, Historical fiction, Love stories, Detective and mystery stories, and Western stories – the authors evaluate different machine learning approaches as solutions for assigning genre.<sup>[3]</sup> Their results show that the tree-based method XGBoost outperformed all models with the highest accuracy at 84% and F1-Measure at 81%. This suggests that machine learning methods can effectively automate genre identification, enabling efficient text classification and retrieval (for another major study of narrative subgenres using machine learning, this time applied to Spanish novels, cf. [Calvo Tello 2021]).

Concerning the extraction of distinctive words for genre classification, Magali Paquot and Yves Bestgen (2009) used three statistical tests, including the Log-likelihood ratio test, the t-test, and the Wilcoxon/Mann-Whitney test, to extract and compare the number and type of keywords from a corpus of academic writing against a corpus of fiction. Their goal was to identify keywords that are “evenly distributed across its component parts (corpus sections or texts) and display a higher frequency and a wider range than in a reference corpus of some kind” [Paquot and Bestgen 2009, p. 19]. They found that the t-test was an adept solution for this task. However, their research also showed that the Log-likelihood ratio test, the T-test, and the Wilcoxon-Mann-Whitney test yielded quite different results for the extraction of distinctive words. Overall, their work highlights the importance of carefully selecting the appropriate statistical measure for a given task, as different measures can lead to varying results.

As the previous research has demonstrated, the question of identifying words that are particularly characteristic of a corpus or text can be addressed using a variety of methods and statistical measures [Kilgarrriff 2001, 233]. However, due to the large number of available options, it can be challenging to select the most appropriate measure (or

2

3

4

5

6

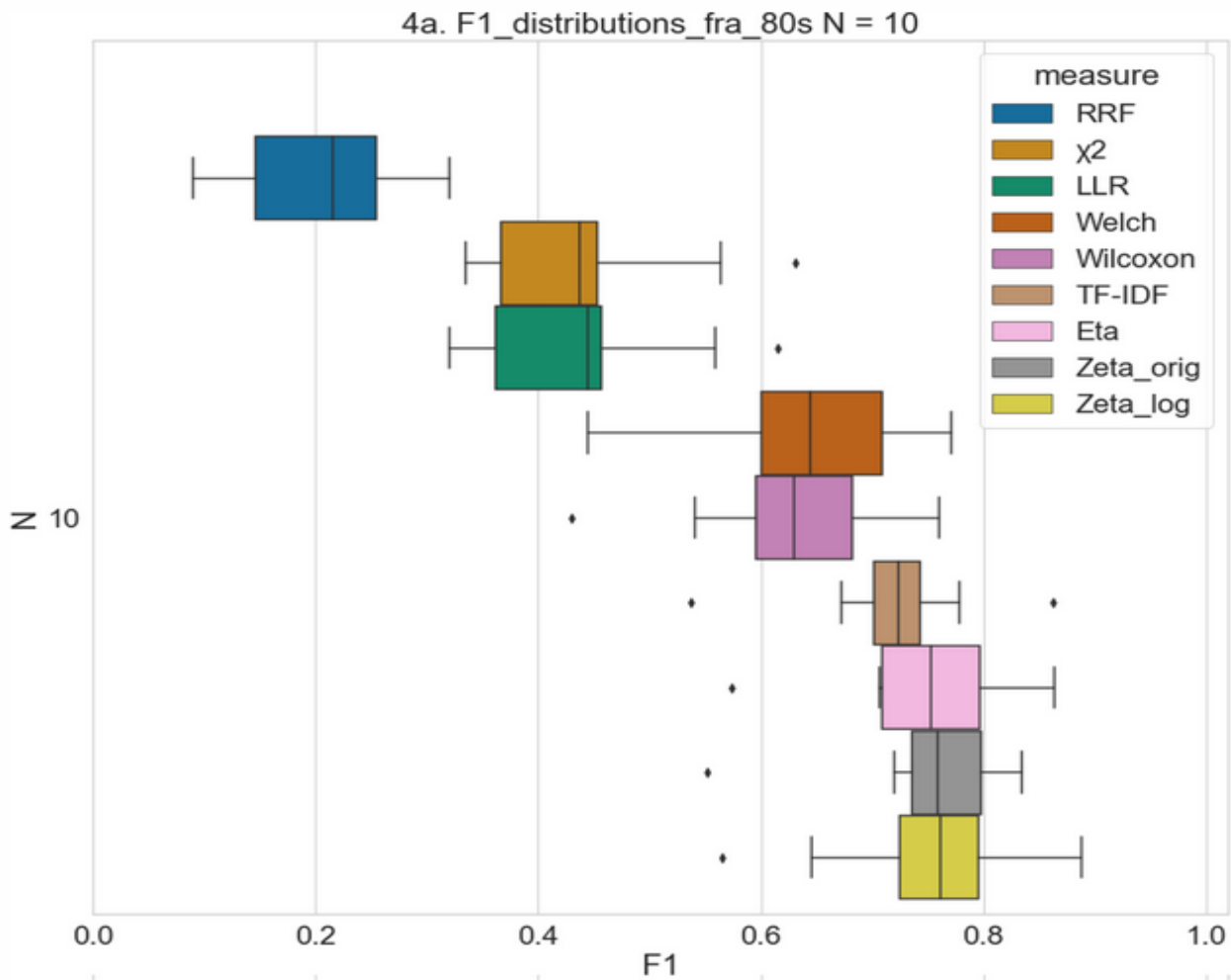
measures) for a given research question or corpus. To address this challenge, the “Zeta and company” project worked on shedding some light on the multitude of available options.<sup>[4]</sup> We first conducted a comprehensive study of qualitative definitions of keyness or distinctiveness [Schröter et al. 2021], then underscored the importance of considering not only frequency but also dispersion in keyness analysis [Du et al. 2021], and also conducted an empirical quantitative evaluation of nine different measures of distinctiveness within the realm of Computational Literary Studies by comparing their performance as feature selectors for document classification [Du et al. 2022b].

## Measures of distinctiveness

Several statistical measures and tools have been introduced and implemented to investigate and analyze large volumes of textual data from a comparative perspective. These measures have been developed and applied in different disciplines, such as Corpus and Computational Linguistics (CL), Information Retrieval (IR), or Computational Literary Studies (CLS). As mentioned earlier, the aim of our research project “Zeta and Company” was to explore several different measures of distinctiveness and provide a better understanding of their characteristics. To achieve this goal, we have implemented nine measures of distinctiveness in a Python framework called *pydistinto*. While these measures all compare the properties of all words in the vocabulary in a target and a comparison corpus to establish a ranked list of distinctive words, they differ according to the way in which they identify distinctive words. We distinguish three groups:

- Frequency-based measures primarily concentrate on the frequency of the words and consider the corpus as a “bag of words,” without considering whether the words are evenly distributed within the corpus or not. Measures falling into this category include the following: Ratio of relative frequencies, Chi-squared test, Log-likelihood ratio test.
- Distribution-based measures are based on a quantification of the words' distributional properties (such as mean / median / mode and standard deviation). The t-test and Welch's t-test belong to this category.
- Dispersion-based measures calculate the degree to which words are more or less evenly spread across a corpus. The following measures belong to this category: Burrows Zeta, logarithmic Zeta, Eta [Du et al. 2021], a TF-IDF-based measure<sup>[5]</sup>, Wilcoxon rank-sum test.

In our prior research, based on the results of the classification accuracy (cf. [Du et al. 2022b]), we found that measures within one group exhibit roughly comparable behavior and results (Fig. 1).



**Figure 1.** Results of document classification based on distinctive words extracted using 9 different measures [Du et al. 2022b].

In this research, the performance of the measures relied on their ability to identify words that serve as features for classifying unseen texts by subgenre. An intriguing observation is that the performance of various measures varied significantly between different groups. We decided not to include all measures available in our framework into the present study, but selectively chose three with the highest performance in their respective groups. The selected measures are logarithmic Zeta (Zeta), Welch's t-test (Welch), and Log-likelihood ratio test (LLR).

The Log-likelihood ratio test, a standard measure in computational linguistics, is a statistical distribution test with inherent hypothesis testing. This measure finds extensive application in computational linguistics (CL) and has been integrated into various corpus analysis tools, including WordSmith Tools [Scott 1997], AntConc [Anthony 2005], and Wmatrix [Rayson 2005]. One of the main issues with this measure is that the hypothesis test is formulated to compare statistically independent events and to treat corpora as a bag of words. Additionally, frequency-based measures rely on the total word count in a corpus, overlooking the potentially uneven distribution of words within it [Lijffijt 2014]. This means that the LLR may favor high-frequency words even if they only occur in a low number of documents.

Welch's t-test, named after its creator, Bernard Lewis Welch, serves as an adaptation of the Student's t-test [Welch 1947]. In contrast to the classical t-test, Welch's t-test is more reliable when dealing with two samples that exhibit unequal variances and potentially unequal sample sizes [Ruxton 2006]. Similar to the Log-likelihood test, it employs hypothesis testing, but distinguishes itself by considering not only the frequency of a feature but also taking into account the sample mean, standard deviation, and sample size in the calculation of the t-value. This implies that the distribution of words within an entire text collection plays a crucial role in the calculation of Welch's t-test.

Originally, Zeta was created in the context of authorship attribution analysis, but it later gained popularity in other tasks in the field of Computational Literary Studies, like genre analysis [Schöch 2018] [Schöch et al. 2018] or gender analysis [Hoover 2010]. In contrast to measures based on statistical tests, Zeta is mathematically straightforward. To calculate the Zeta score, the longer texts such as novels are initially segmented, e.g. into 5000-word segments. The binary document frequency of each word in each part of the corpus is determined as the document proportion of that word. In the subsequent step, the normalized document proportion value for each word in the comparison corpus is subtracted from the corresponding value in the target corpus, resulting in the final Zeta-Score for each word. The higher the Zeta-score, the more significant the word. As a measure of distinctiveness, Zeta was initially introduced by John Burrows [Burrows 2007] and is mathematically defined as:

$$Zeta_i = docp_i(T) - docp_i(C)$$

where  $docp_i(T)$  and  $docp_i(C)$  are the normalized document proportions of the word  $i$  in the target corpus and the comparison corpus, respectively. The normalized document proportion is calculated as:

$$docp_i(T) = \frac{docf_i(T)}{n(T)}$$

where  $n(T)$  is the number of documents in the target corpus, where word  $i$  occurs at least once. Fundamentally, Zeta quantifies the degrees of even dispersion of a feature in two text collections and compares them. The logarithmic Zeta used in our current research was adapted from Burrows Zeta by [Schöch et al. 2018].

In the following section, we describe the corpus of contemporary French novels that we curated and employed as the foundation for our research.

## Application Domain: The Contemporary French Novel

The French popular novel has a long history, going back at least to the mid to late nineteenth century and the age of industrialisation.<sup>[6]</sup> It is true that the period from 1860–1920 is often seen as the “golden age” of the popular French novel, when phenomena like the “roman-feuilleton” and the serial novel (with recurring protagonists like Rocambole and Rouletabille) occurred. As a consequence, earlier studies of the French popular novel tended to focus on the nineteenth and early twentieth centuries [Angenot 1995] [Olivier-Martin 2013].

However, there is a large and differentiated landscape of popular novels in the second half of the twentieth century as well [Migozzi 2005]. These novels are published by long-standing publishing houses (e.g. Harlequin, Fleuve noir, Librairie des Champs-Élysées, Tallandier) in specialized collections with large print runs, high publication frequency and distinct audiences (e.g. Harlequin’s collection “Rouge passion” as opposed to Fleuve noir’s “Anticipation” or the Librairie des Champs-Élysées’ “Le Masque”).<sup>[7]</sup> Specific popular genres did receive particular critical attention, however, whether it is the “roman policier” [Todorov 1971] [Vanoncini 2002] [Dubois 2005], the science-fiction novel [Slusser 1989] [Thomas 1989] [Millet and Labbé 2001] [Baudou 2003] [Mather and Rheault 2016], or the “roman sentimental” [Helgorsky 1985] [Helgorsky 1987] [Constans 1999].

Recently, renewed attention has been given to the contemporary French popular novel from a linguistically inflected perspective, notably in the context of the project “Phraséorom” as well as in a French research group called “Motifs.” Both groups have been focusing on developing techniques for extracting distinctive lexico-grammatical patterns from corpora structured by subgenres, such as the “roman policier” and “roman sentimental” (e.g. [Gonon et al. 2017]). Our domain of application connects to this research, focusing on (popular and high-prestige) novels from the period 1970 to 1999.

Series of crime or science fiction novels are examples of serial storytelling, which can look back on a long history in French literary history.<sup>[8]</sup> Sentimental novels, crime novels, and science fiction novels also belong to the broader category of popular novels, sometimes pejoratively called “littérature de gare” in France, as the novels are often sold in train stations and similar spaces. Also termed schema or formula literature, it uses recurring formulae of structural

elements, typical protagonists and character constellations, narrative patterns, typical settings, characteristic linguistic patterns, or a recurrent narrative structure.<sup>[9]</sup> Recurring patterns and schemes within these subgenres make them particularly suitable for a study of extracting distinctive words per genre using different measures of distinctiveness, since the horizon of expected patterns of thematic clusters, of settings, tonalities, or linguistic patterns is more restricted within these popular novel subgenres than in the wider and more heterogeneous domain of literary fiction more generally.

To identify the common features of crime stories, science fiction novels, sentimental novels, and literary fiction, we began by qualitatively compiling a list of typical elements such as thematic concepts, main characters, linguistic patterns, tonality, and narrative form mentioned in relevant scholarly literature on these subgenres. Our aim was to construct general profiles of these subgenres through a close reading of secondary sources.<sup>[10]</sup>

18

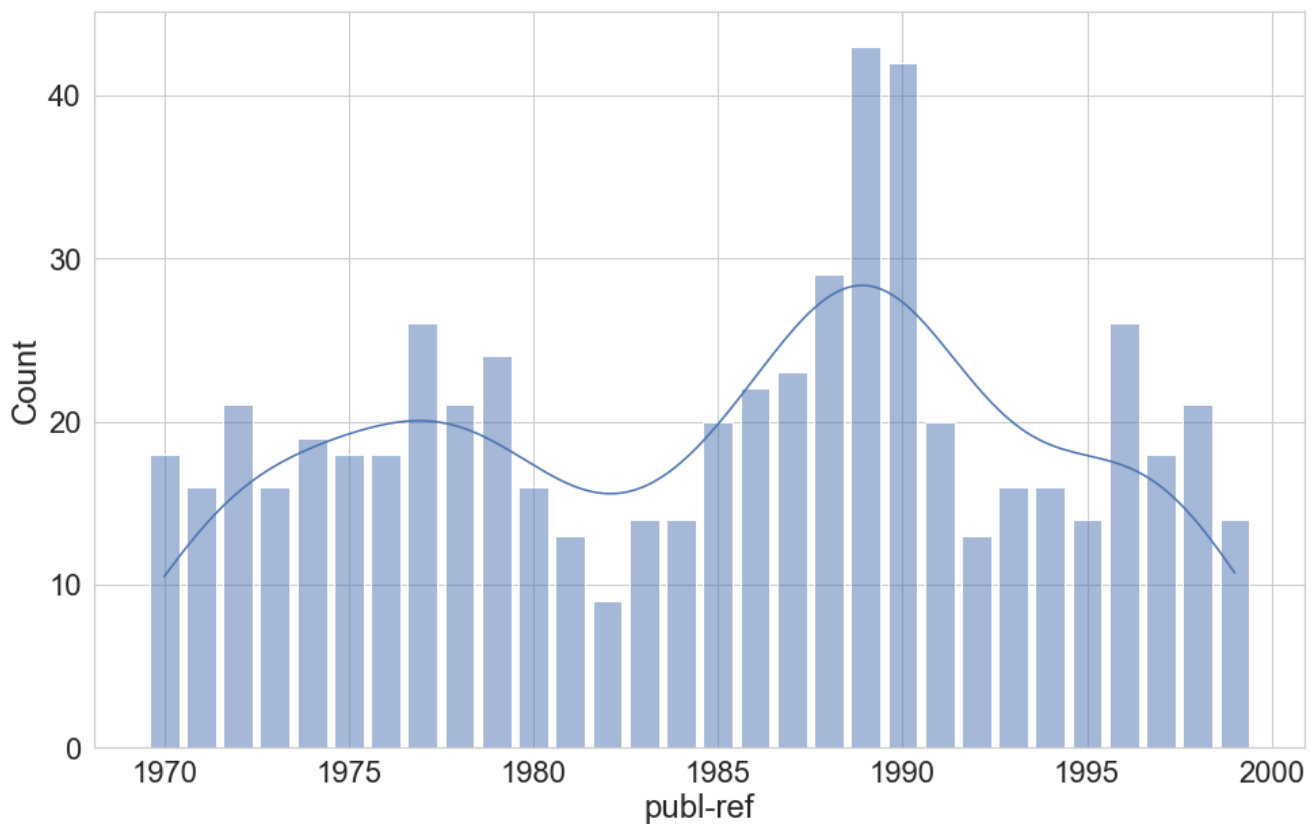
Using scholarly literature on these subgenres, we condensed typical elements for each subgenre into concise qualitative textual profiles structured around thematic concepts, main characters, language patterns, settings, tonality, narrative form, and narrative structure. These textual profiles were written prior to extracting distinctive words per subgenre and were meant to capture salient aspects of crime novels, science fiction novels, sentimental novels, and literary fiction that are frequently mentioned in the scholarly literature. Our approach was to then, in a second step, align distinctive words identified by the statistical measures of distinctiveness with these subgenre profiles in order to gauge the level of interpretability of the distinctive words in relation to the subgenres.

19

## Data: Four Subgenres, Three Decades

In a first step, a corpus of contemporary French fiction was built. It is composed of 600 entire novels balanced along two categories: four subgenres (sentimental novels, crime novels, science fiction novels, literary fiction) and three time periods (1970s, 1980s, and 1990s).<sup>[11]</sup> The average length and median length of the novels are 55,050 and 46,935 words, respectively. The longest and shortest of these novels contain 240,064 and 12,883 words, respectively. In total, the entire corpus contains over 33 million words and 207,207 distinct word types.

20



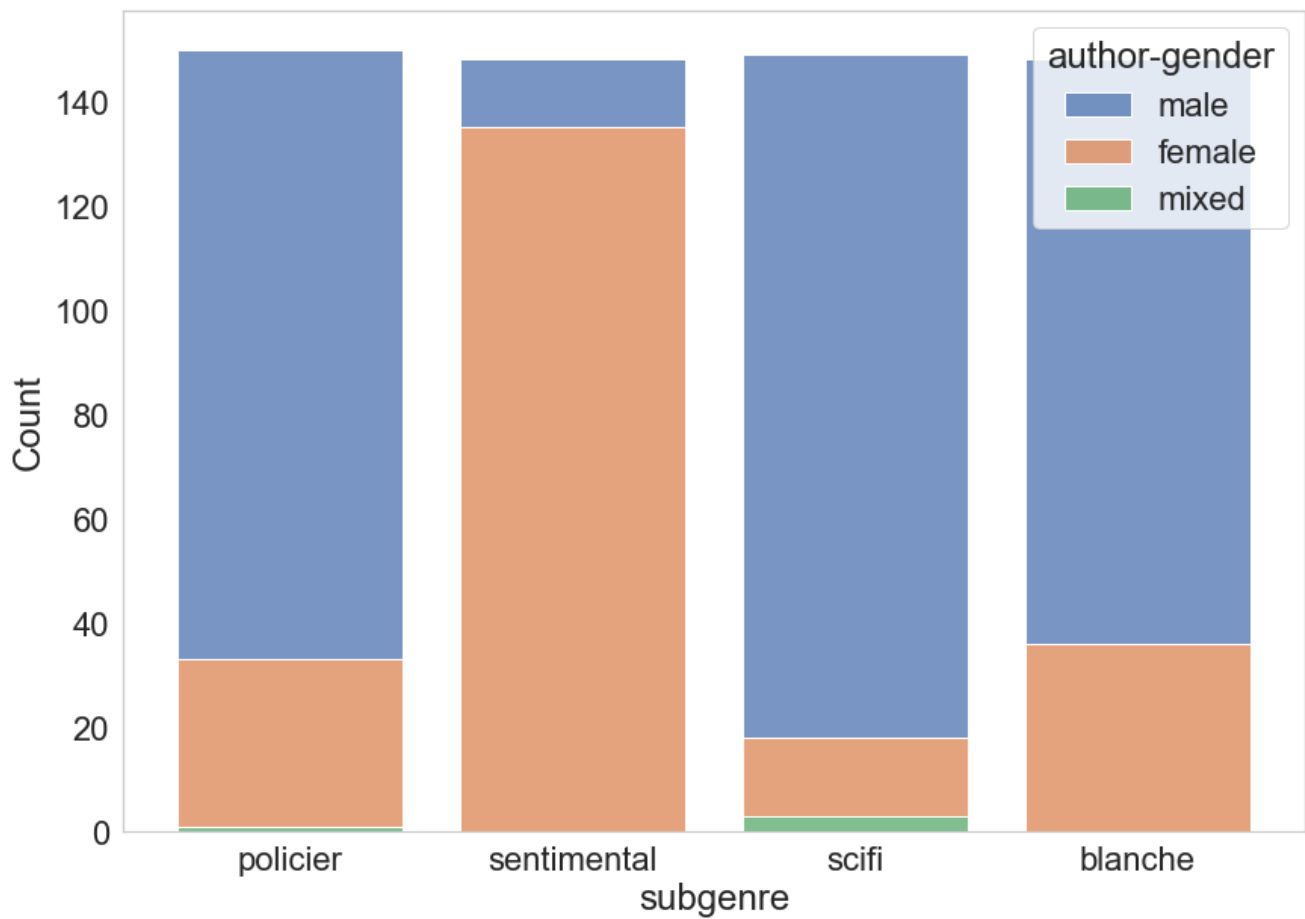
**Figure 2.** Year of publication of the novels in the corpus.

Concerning the year of first publication, the corpus consists of 50 novels per decade and per subgenre. The decades are evenly represented, but Figure 2 shows that there is a peak of novels in the corpus in 1989–1990 across all subgenres.

21

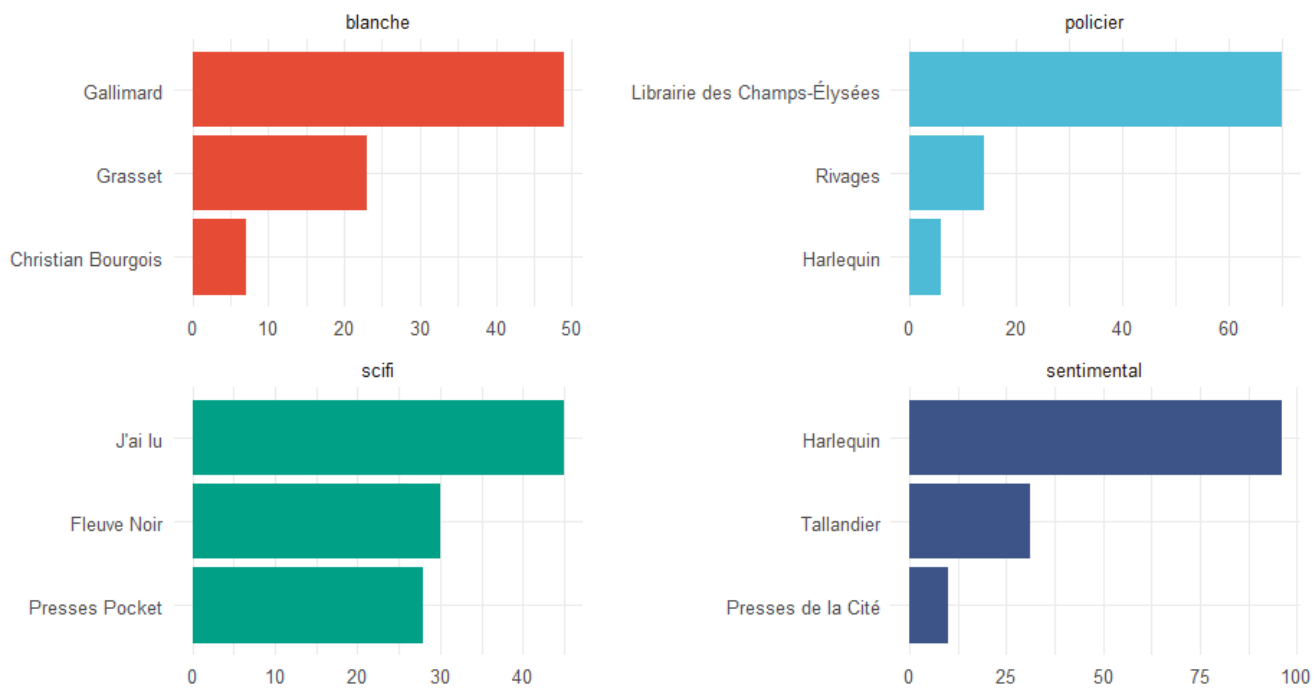
As one can observe in Figure 3, the gender balance per subgenre is not even. Crime novels, literary fiction, and science fiction novels have a much greater share of male authors in the corpus (81.3% on average), whereas the share of female authors is much greater in the sentimental subgenre (91.2%). Not only the part of female authors is greater, but the share of female readers is also immense (cf. [Jensen 1980, p. 111]). It is true that statistically-speaking, then, the subgenre distinction is in an undue correlation with author gender, in our corpus. However, this aspect of corpus composition does reflect the publishing situation in the French popular novel. Results based on a corpus in which male and female authorship were represented equally across all subgenres, or results based on stratified sampling from our present corpus, would, therefore, not be an accurate representation of the literary landscape.<sup>[12]</sup>

22



**Figure 3.** Overview of author gender per subgenre in our corpus.

Concerning the publishers, Figure 4 gives an overview of the corpus composition. While *Harlequin* is a big and international publisher of romance novels, the French publishing house *Gallimard* is known for publishing literary fiction. With a decidedly worldwide marketing strategy for an international public and explicitly formulated writing guidelines for authors, *Harlequin* is a special case among the publishers in the corpus.<sup>[13]</sup> *Fleuve Noir* is an influential publisher of French science fiction and crime fiction (as well as espionage novels, which are not currently included in our corpus, however), whereas *Librairie des Champs-Élysées* specializes in crime novels.



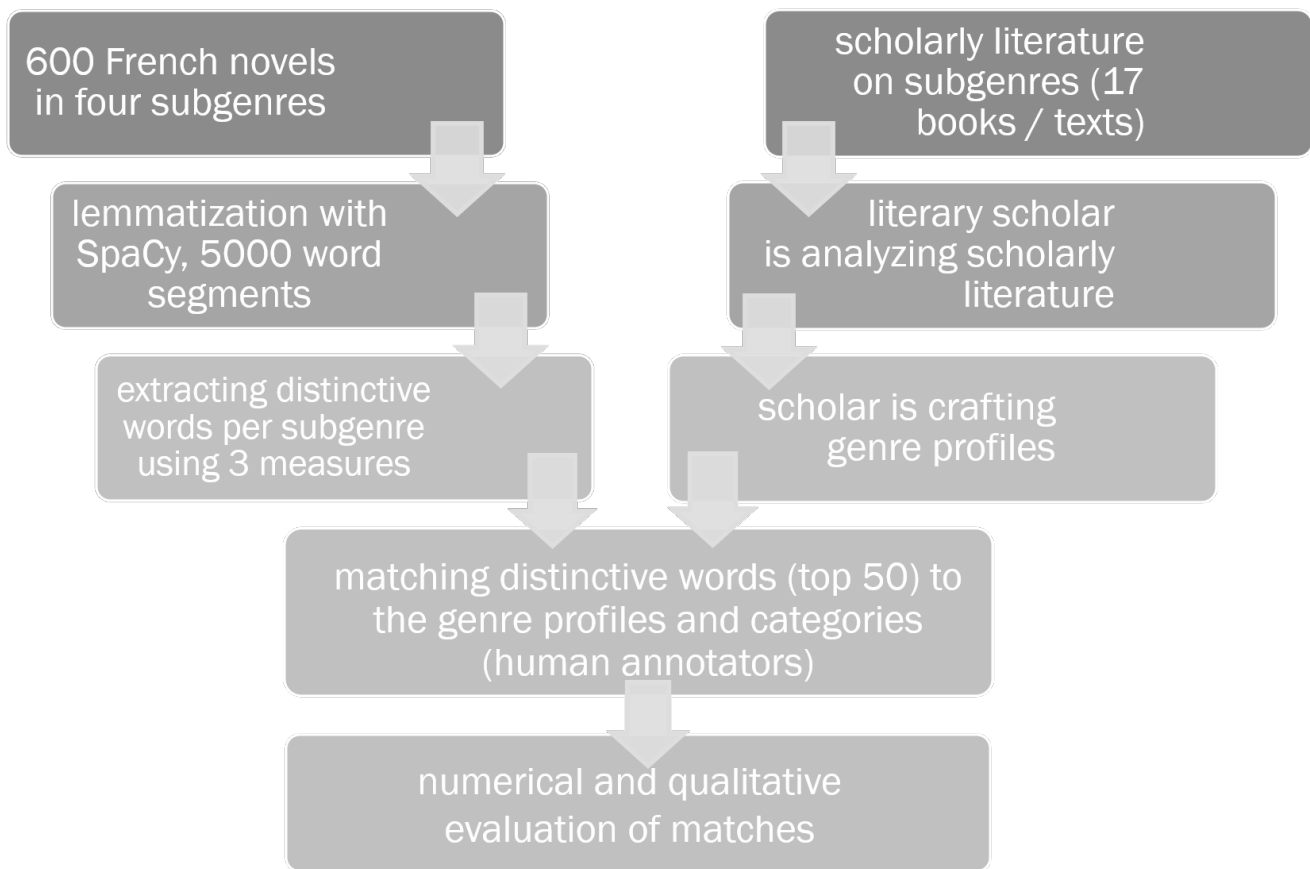
**Figure 4.** Most frequent publishers per subgenre in our corpus.

All novels in the corpus are still protected by copyright, which hinders the full publication of our research data concerning the full texts. However, we have published the corpus in the form of “derived text formats” (also known as “extracted features”), an innovative format which transforms the original text in such a way that the resulting files are still suitable for some computational methods of analysis, but no longer readable by humans (cf. [Schöch et al 2020] [Organisciak and Downie 2021]). With the help of this strategy, we hope that transparency of data and the possibility of a reproduction of our study is possible for interested scholars [Du and Schöch 2024].

## Methods: Expertise vs. Statistics

Because our goal is to track the words which are characteristic for each subgenre using the three different measures of distinctiveness, we carried out our tests by comparing each single subgenre to the other three subgenres in our corpus. In this case, the comparison takes place between a target corpus (containing 150 novels of one subgenre) and a comparison corpus (containing the remaining 450 novels spread across three subgenres).

French is a highly-inflected language; therefore, all the texts were lemmatized using spaCy [Montani et al. 2023].<sup>[14]</sup> They were then processed using the python package *pydistinto*, which has been developed within the project “Zeta and Company.”<sup>[15]</sup> We did not use stopwords lists, as we are interested in the entire output of the different keyness measures in our evaluation. To support the calculation of Zeta (see above), the novels were also split into 5000-words-segments. As mentioned before, we applied three measures of distinctiveness in our test, at the level of the lemmas: Zeta, Welch, and LLR. As a result, we obtained three lists of distinctive words in each of the four comparative tests (one for each subgenre), resulting in 12 lists of distinctive words. Our initial finding revealed a certain level of overlap in the distinctive words identified by the three measures for the same genre. To quantify this overlap, we calculated the Jaccard similarity (see [Niwattanakul et al. 2013]) for the top 50 words. The average similarity (where higher values indicate greater overlap) was 0.43 between Zeta and Welch, 0.21 between Welch and LLR, and 0.05 between Zeta and LLR. We then selected the top 50 words from each list for the alignment with the qualitative subgenre profiles.



**Figure 5.** Workflow of statistics vs. expertise.

Our expectation was that the extracted distinctive words per subgenre, contrasted against the texts in all remaining subgenres, would yield words characteristic for the subgenre in the categories described in the genre profiles: thematic concepts (such as murder), typical language patterns (such as sociolect), main characters (such as detective), space/setting (such as urban setting), tonality (such as emphasis on emotion), narrative form (such as dialogue), narrative structure (such as “Cinderella plot involving a dramatic event and a final reunion”).

The approach was to then match each of the top 50 most distinctive words to these subgenre profiles and count the hits per category as words that were both *expected*, based on the subgenre profiles, and in fact *observed* in the statistically-generated most distinctive word lists (“expected and observed words”). The process of matching involved some more straightforward matches (e.g., matching the word “murder” to the thematic concept of “murder” in the crime novels), but also words that had to be interpreted to see whether they fit into the semantic field described (e.g., identifying the words “soldier” and “german” as belonging to the thematic field of “history” or “war” in literary fiction). All matches are documented in the repository.<sup>[16]</sup>

We also consider that a matching between a distinctive word and a particular aspect of the genre profile provides an indication of the interpretability of the extracted word. For the purpose of this study, we define interpretability at the granular level of individual words: A word is defined as interpretable if a human annotator can associate it with a specific section of a genre profile.<sup>[17]</sup> As [Benois-Pineau et al. 2023] succinctly state, to interpret means “to explain the meaning of something.” In addressing interpretability, the overarching question is furthermore: Can we trust the model [Masís 2023, p. 6]? In our case, this question translates into: Can we trust the keyness measure? The higher the proportion of interpretable words, the more reliable the measure becomes for similar tasks.

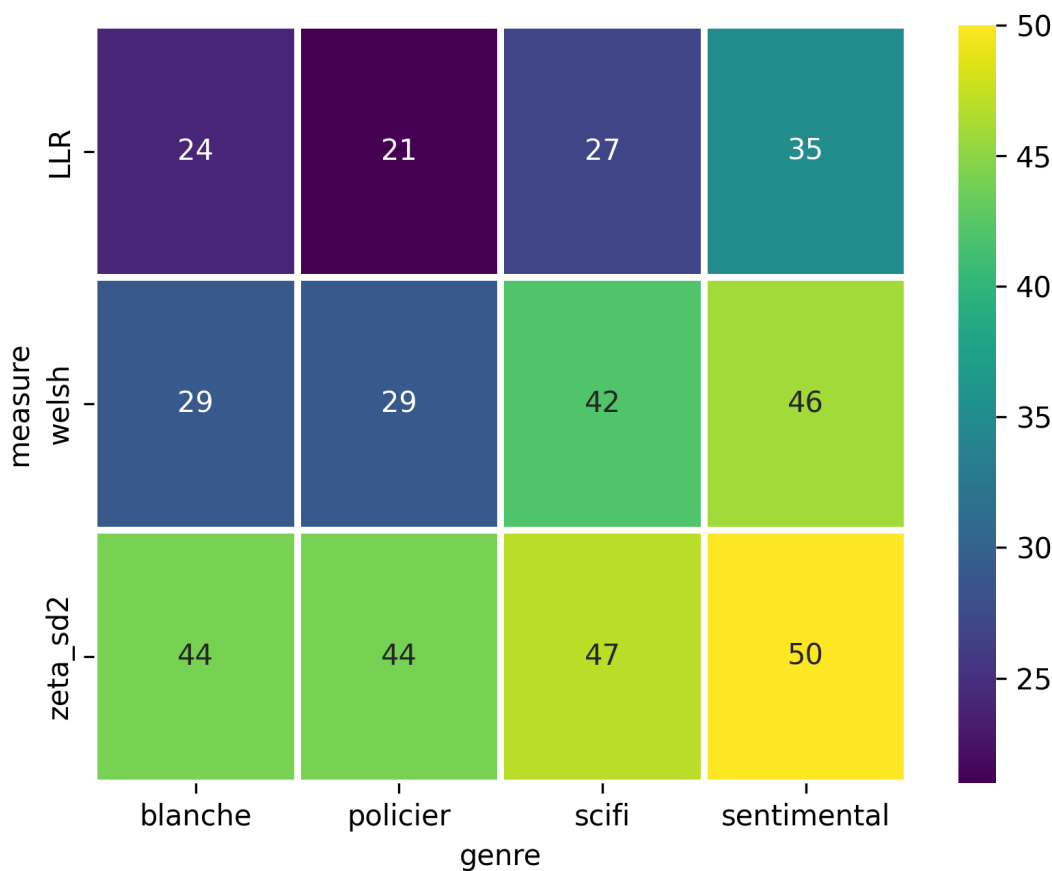
The goal of our procedure was to quantify the proportion of interpretable, matching distinctive words per measure of distinctiveness.<sup>[18]</sup> Within the logic of our evaluation, words that are present in the word lists, but cannot be matched to any aspect of the subgenre profiles are indicated as “unexpected and observed words.”<sup>[19]</sup> If a high number of words

from a given word list can be matched to the relevant subgenre profile, we interpret this to mean that the measure of distinctiveness used to generate the word list has performed well.

The matching between word lists and subgenre profiles was performed independently by three scholars of French literature. Since the distinctive words were interpreted without surrounding text, it is not surprising that different scholar's matching can vary.<sup>[20]</sup> Therefore, we have discussed the results of the matching, in a second step. If there were at least two scholars with the same matching results, then their annotations were used. If all three were different, then the word was marked as “unclear.” In the following, we describe the matches for each measure and subgenre, summarizing the findings supported by the more detailed data that can be found in the data repository.

## Results

Based on the procedure described, we obtain the following results, organized here by measure of distinctiveness. Figure 6 provides an overview of the results.



**Figure 6.** Comparing interpretability of distinctive words per keyness measure: log. Zeta, LLR, and Welch's t-test.

### Logarithmic Zeta

Starting with logarithmic Zeta, and concerning the matching with the sub-genre profile for detective fiction, we see that distinctive words in the categories language patterns and main characters were frequently found. The expected use of sociolects, informal register, or direct speech was observed within the distinctive words. However, the category tonality was hard to grasp with distinctive words: none of the 50 most frequent words that Zeta had identified reflected phenomena like cynicism or sarcasm, features that are hardly visible on a word level. Distinctive words relevant to the

consumption of alcohol and tobacco such as *cigarette* ("cigarette"), *bouteille* ("bottle"), *bière* ("beer") were also found for the subgenre of crime fiction.<sup>[21]</sup>

With respect to science fiction, the extracted content words could be matched to narratives typically delving into the core fears and aspirations of humanity, addressing issues such as overpopulation, totalitarian society, evil forces, (intergalactic) colonialism, or self-destruction of mankind by technical progress. Concerning language patterns in science fiction, we observed a lot of technical vocabulary extracted by logarithmic Zeta.

For sentimental novels, the deliberate emphasis on emotion, antitheses, and direct appeals to the reader's emotions, fostering a heightened sense of pathos, was visible in the extracted distinctive words. The classic Cinderella plot structure involving a dramatic event – visible in lemmata like *surprise* ("surprise") or *brusquement* ("suddenly") – and a final reunion of the couple in a happy ending was also visible in the resulting distinct words in Zeta with *épouser* ("to marry") or *mariage* ("marriage").

The genre of literary fiction or "littérature blanche" was certainly the most heterogeneous and therefore also the most difficult for abstracting into a profile and for matching distinctive words to the categories. However, thematic concepts like political problems of the present, dissolution of families, postmodern lifestyle, and narratives exploring history or war were reflected in the extracted words with logarithmic Zeta. Unexpected words visible in the distinctive lemmata of literary fiction are words relating to animals like *cheval* ("horse") and *oiseau* ("bird") or plants like *fleur* ("flower") and *rose* ("rose"), which may also indicate a color.

Overall, we can note that Zeta's strength lies in its ability to extract content words: in comparison to the other measures, the highest share of interpretable words could be extracted when using Zeta.<sup>[22]</sup>

## Welch's t-test

Next, we report on the matching of the distinctive words obtained using Welch's t-test with the subgenre profiles. With respect to crime novels, the narrative structure with a central axis of elucidation and investigation was reflected in extracted words like *enquête* ("investigation"), and *téléphone* ("telephone"). Other than thematic concepts, an urban setting being characteristic within the detective genre, the distinctive words *voiture* ("car"), *bureau* ("office") could be matched with these expected literary spaces within the detective genre.

Concerning science fiction, Welch's t-test extracted a relatively large quantity of content words. The language pattern of technical vocabulary was very dominant in the distinctive words extracted by Welch's t-test. As Ted Underwood noted, adjectives denoting size such as *huge*, *gigantic* and even *tiny* serve as reliable indicators that a book may be classified as science fiction [Underwood 2019, p.41]. In our analysis of extracted words, we observed several terms, including *centaine* ("hundreds"), *mètre* ("meter"), *nombre* ("number"), and *nombreux* ("numerous") as belonging to this category of textual clues for the genre science fiction. A bit surprisingly, very few typical main characters such as scientists, doctors, robots, or artificially intelligent beings were extracted, except for the abstract term *créature* ("creature"). More hits were found for spaces and settings of science fiction.

For the sentimental subgenre, the expected thematic concepts of passions, evolving relationships, and sensual scenes had several matches within the extracted words. The most visible category of matches, however, was the category of narrative form for the sentimental genre. When in general, the category of narrative form was hard to grasp with distinctive words for the other subgenres, the sentimental novels were an exception here: A significant proportion of the content in sentimental novels is conveyed through dialogue, and this large proportion of dialogue in the genre was reflected by a multitude of verbs of communication and further signs of direct speech in the extracted distinctive lemmata.

In literary fiction, problematic, failing relationships or failing happiness is typical and is visible in the extracted *malheur* ("unhappiness"). Furthermore, urban spaces can be characteristic and are reflected in the extracted words with *Paris* and *rue* ("street"). Other than that, Welch's t-test extracted typical thematic concepts for the genre similar to Zeta.<sup>[23]</sup> The specific textual pattern of *écrire* ("to write") in literary fiction is an element extracted by both logarithmic Zeta and

Welch's t-test and something that has also been described by researchers within the Phraséorom project [Gonon et al. 2018].

## Log-likelihood ratio test

Moving on to LLR, when comparing all three measures, it is noteworthy that the proportion of “unexpected and observed words” (i.e., uninterpretable words) is consistently highest across all four subgenres. For the crime genre, distinctive words primarily revolve around typical main characters like “inspector”, “commissioner”, “police”, “murderer”, “policeman”, “sheriff”, “sergeant”. However, many proper names were also extracted with this measure. Given the introspective dimension often found in popular detective novels through first-person narration, the pronoun *je* (“I”) was also identified.

In the science fiction genre, a limited number of thematic concepts like *humain* (“human”), *terrien* (“earthling”) and typical spaces like *planète* (“planet”), *espace* (“space”) and *univers* (“universe”) were extracted. However, it is noteworthy that 23 out of the top 50 distinctive words could not be matched to the genre profiles (see Figure 6).

Similarly, for sentimental novels, a significant proportion of distinctive words comprised words that could not be matched to the genre profiles, in particular pronouns, articles and other highly frequent grammatical words and proper names.

An overall finding across all subgenres is that Log-likelihood ratio test – which is implemented in a lot of corpus analysis tools – performed the weakest in extracting relevant distinctive words per sub-genres according to the qualitative profiles in our setting.

According to general remarks in [Scott 2009, p.80], a limitation of this study is that the choice of the reference corpus affects the results. In our case, the reference corpus consists of the total of all texts minus the subgenre currently under investigation; this affects the results in so far as features that are shared by the reference corpus and the subgenre subcorpus will not surface. For example, distinctive words that are specific to popular literature in general, rather to any one particular popular subgenre, might not be extracted, because in our experimental design, there are popular novels in both target and comparison corpus. Only features where there is a significant departure from the reference corpus (serving as the norm) will be visible in the lists that were matched with the qualitative profiles.

## Conclusion

What can we learn from our findings? We have shown that measures of distinctiveness are able to extract distinctive words per genre, which reflect to a certain extent what one would expect to find as typical for a certain genre. Some subtle elements of certain literary genres like tonality or narrative form or narrative structure are harder to grasp through distinctive features at the single word level. Similarly, sociolect and other informal registers were more observable in crime fiction, but difficult to grasp in other genres. Main characters, spaces, but above all thematic concepts of genres are more accessible categories for matches between distinctive words and qualitative genre profiles.

Differences between popular genres (sentimental novel, crime novel, science fiction novel) and literary fiction persist in our setting, insofar as the popular genres showed more hits per category than the literary fiction. To a certain extent, this is to be expected, as literary fiction is a broader and more heterogeneous category.

When comparing the measures, we found that the measures showed different performances for the task of finding characteristic distinctive words for a certain subgenre.

- Zeta proved to be the most effective in extracting content words with a high degree of interpretability, particularly for detective fiction, science fiction, and sentimental novels. The extracted words matched well with the subgenre profiles.
- Welch's t-test also performed well, especially for sentimental novels where a significant number of dialogue-related words were extracted. The test captured narrative patterns and thematic concepts effectively, though it did not extract as many main character types for some subgenres.
- LLR performed the weakest in aligning distinctive words with subgenre profiles. A high number of extracted

words were either grammatical words or proper names, which were often categorized as “unexpected.”

In comparing the output of three measures of distinctiveness with human made literary genre profiles, our evaluation showed that logarithmic Zeta and Welch’s t-test have a higher share of interpretability of distinctive words than LLR for French novels. Interpretability was operationalized as a match between a distinctive word and a human crafted subgenre profile in prose. LLR’s performance may come as a surprise, considering its widespread implementation in numerous corpus analysis tools, such as Antconc or Wordsmith Tools.<sup>[24]</sup> The reason might be that the data in corpus linguistics differs often in its form, for example news texts, political speeches and similar textual genres that, among many other differences, are usually shorter than literary fiction. In addition, in presidential debates texts, proper names (e.g. “Kamala Harris”, “Donald Trump”) are meaningful or “interpretable” distinctive words.

50

For the domain of Computational Literary Studies, our finding raises questions about the default measures of distinctiveness incorporated into tools. It suggests that a critical revision of these default measures is warranted, at least in scenarios where the application domain is long-form narrative fiction. With the implementation of *pydistinto*, we offer a framework with nine different keyness measures, openly available on GitHub [Du et al. 2022a]. Three variants of Zeta are available furthermore in Stylo, a tool developed by Maciej Eder, Jan Rybicki and Mike Kestemont [Eder et al, 2016].

51

More generally, we believe this method of comparing texts and extracting distinctive words, due to its relevance and ease of use, can be widely applied to research in other fields in the humanities. In many cases and scenarios, it is meaningful to conduct large-scale text comparisons by dividing text collections into two corpora (a target and a comparison corpus) and extracting distinctive words across different eras, political spectrums, religious domains, etc., in order to obtain some (new) insights about these texts from a statistical perspective.

52

As future work, we plan to extend our research by using more complex and semantically-rich linguistic features such as multiword expressions or topics and we are expecting more insights into the distinctiveness of the investigated subgenres.

53

## Author contributions

- Julia Röttgermann: Investigation, Formal Analysis, Visualization, Writing - original draft, Writing - review and editing.
- Keli Du: Investigation, Visualization, Software, Writing - original draft, Writing - review and editing.
- Julia Havrylash: Investigation, Methodology, Writing - original draft, Writing - review and editing.
- Christof Schöch: Funding acquisition, Data curation, Supervision, Writing - review and editing.

## Data and code

- Corpus and metadata: <https://github.com/Zeta-and-Company/DTF600>, DOI: 10.5281/zenodo.10853581.
- Code and research data: <https://github.com/Zeta-and-Company/expertise-statistics>, DOI: 10.5281/zenodo.10853663.
- Pydistinto: <https://github.com/Zeta-and-Company/pydistinto>, DOI: 10.5281/zenodo.6517683.

## Acknowledgements

“Zeta and Company. Measures of Distinctiveness for Computational Literary Studies” is funded by the German Research Foundation (DFG, project number 424211690) in the framework of the Priority Programme Computational Literary Studies (2020–2023). Many thanks to Dr. Élodie Ripoll (Trier University) and Ariadne Baresch (Trier University) for annotating the lists of distinctive words.

54

## Notes

[1] The project “Zeta and Company” has been conducted at Trier University, Germany, and is funded by the Deutsche Forschungsgemeinschaft (2019-2023, project ID 424211690) as part of the priority programme on *Computational Literary Studies* (SPP 2207).

[2] “Littérature blanche” designates high-prestige literary fiction, the color term being an allusion to the color of the book covers of French

literary fiction being typically white, as for example in the case of the “Éditions Gallimard”, a publisher that even has a long-running and very prestigious book series called “Collection Blanche.”

[3] The machine learning approaches evaluated by Worsham and Kalita are Convolutional Neural Networks, LSTMs, HAN, Naive Bayes, k-Nearest Neighbors, Random Forests and XGBoost [Worsham and Kalita 2018, p. 1963].

[4] See the acknowledgements section for details.

[5] We proposed this TF-IDF variant, which compares the difference between the TF-IDF-weighted absolute frequency of a word in the target and the comparison corpus. Although in the following we refer to it simply as TF-IDF, it is an adaptation of TF-IDF.

[6] Popular novels can be defined as “print stories that are mass-produced and mass-consumed” [Hoppenstand 2016, p. 102] and structured under specific genre conventions in a predictable formula.

[7] Note that the subgenre assignment for the works in our corpus is determined by the works belonging to a given publishing house's subgenre-based collection.

[8] Examples of serial storytelling can be found in the French novel of the 19th century, for example in the *Rougon-Macquart* story cycles by Émile Zola or in *La Comédie humaine* by Balzac. Famous examples from the 20th-century include the *Nouveaux Mystères de Paris* by Léo Malet (crime fiction) or the *Compagnie des glaces* series by Georges-Jean Arnaud (science fiction).

[9] Bertolt Brecht describes the phenomenon of formula literature within his essay on the crime novel “Über die Popularität des Kriminalromans”: “The crime novel is about logical thought and demands such logical thought from the reader. (...) Correspondingly it has a formula and displays its strength in variation thereof.” (Brecht, 1967 [1938], p. 450–458; English translation as in [Brecht 2004, p.90]).

[10] These profiles serve as abstract representations of the typical traits associated with each genre, not necessarily formulated in respect of distinctive words (for example: irony or ellipses). The profiles are written by a domain expert (Romance literary studies); the used references are listed in the repository: <https://github.com/Zeta-and-Company/expertise-statistics>.

[11] For the full overview of metadata, see the repository linked in the section “Code and Data.”

[12] In addition, with many of the authors of popular fiction being little-known, we can often only assume the author gender from the gender conventionally associated with their first names. This also means that there is likely to be a certain proportion of cross-gender pseudonymy involved in the corpus whose extent we are currently not able to ascertain.

[13] Harlequin novels are “a product of an industrialization of fiction,” as [Dudovitz\_2019, p.120] points out (cf. [Rabine 1985, p. 39]).

[14] We used the “fr\_dep\_news\_trf” model: [https://spacy.io/models/fr#fr\\_dep\\_news\\_trf](https://spacy.io/models/fr#fr_dep_news_trf), last access: 19 February 2024.

[15] See: <https://github.com/Zeta-and-Company/pydistinto>, pydistinto v0.1.2, DOI: <https://doi.org/10.5281/zenodo.6327967> [Du et al. 2022a]. <https://doi.org/10.5281/zenodo.6327967>

[16] Code and research data: <https://github.com/Zeta-and-Company/expertise-statistics>.

[17] In a similar sense, [Chang 2009] designed an approach where a human evaluator is tasked with identifying the “odd one out” from a set of topic modeling results, using this task to assess the interpretability of the topic words.

[18] For a further discussion of salience, distinctiveness and aboutness in the context of keyword extraction (cf. [Schröter et al. 2021]).

[19] Note that some of the extracted keywords may be plausible for the subgenre post hoc, but did not match to the profiles as we had previously established them.

[20] We can report a mean inter-annotator agreement (Krippendorff's alpha) of  $\alpha=0.083$  for matching of words to a specific category and a mean of  $\alpha=0.139$  for the agreement on the interpretability of words.

[21] These words were not matched with the subgenre profiles. For a corpus-stylistic study on alcohol and tobacco consumption in English and French novels since the 1950s, see [Grossman et al. 2020].

[22] See also [Gittel and Köppe 2022] who discuss the explainability and interpretability of models distinguishing fictional and non-fictional texts [Piper 2016] and the life cycles of genre (Underwood, 2016).

[23] In a study on the authorial style of Collins and Besant, D. Hoover compares the overlap of words extracted by Zeta versus words extracted by t-test. Concerning Zeta words, he states, that “fewer function words and more meaningful words” are extracted compared to t-test [Hoover 2020, p. 250–71].

[24] We can agree with [Brezina 2018, p.84–85] in this regard, that there is growing evidence that “log likelihood used for corpus comparison is prone to identifying far too many keywords (false hits).” See also [Brezina and Meyerhoff 2014]. Alessia Tranchese also discusses the pitfalls of Log-likelihood in her study on rape coverage in the British press. She states that there could be “words with relatively high frequency - but uneven distribution - which an LL calculation could still retrieve as ‘key’”. This might happen, for instance, if a word is rather unique, like the name of a person, or the name of a city” [Tranchese 2023, p.99].

## Works Cited

**Angenot 1995** Angenot, M. (1975) *Le roman populaire: recherches en paralittérature*. Montréal: Presses de l'université du Québec. Available at: [http://digitale-objekte.hbz-nrw.de/storage/2009/11/07/file\\_18/3292944.pdf](http://digitale-objekte.hbz-nrw.de/storage/2009/11/07/file_18/3292944.pdf) (Accessed: 6 November 2024).

**Anthony 2005** Anthony, L. (2005) “AntConc: Design and development of a freeware corpus analysis toolkit for the technical

- writing classroom”, in *IPCC 2005. Proceedings. International Professional Communication Conference, 2005*. Limerick: IEEE, pp. 729–737.
- Anthony 2024** Anthony, L. (2024) *AntConc (Version 4.3.1)* [Computer Software]. Tokyo, Japan: Waseda University. <https://www.laurenceanthony.net/software/AntConc> (last accessed 27 August 2025)
- Baker 2004** Baker, P. (2004) “Querying keywords: Questions of difference, frequency, and sense in keywords analysis”, *Journal of English Linguistics*, 32(4), pp. 346–359. Available at: <https://doi.org/10.1177/0075424204269894> (Accessed: 6 November 2024).
- Baudou 2003** Baudou, J. (2003) *La Science-fiction*. 1e édition. Paris: Presses Universitaires de France - PUF (Que sais-je?).
- Benois-Pineau et al. 2023** Benois-Pineau, J. et al. (2023) *Explainable Deep Learning AI: Methods and Challenges*. London: Elsevier.
- Bentolila 2016** Bentolila, É. (2016) *Le roman policier français de 1970 et 2000 : une analyse littéraire*. Thèse de doctorat. Grenoble: Université Grenoble Alpes (ComUE). Available at: <https://www.theses.fr/2016GREAL013> (Accessed: 6 November 2024).
- Bleton 2014** Bleton, P. (2014) “‘Meurtre’ ne rime à rien. La ville dans le roman policier français des années 1958-1981”, *Revue critique de fixxion française contemporaine*, 10, pp. 13–23. Available at: <https://doi.org/10.4000/fixxion.10005> (Accessed: 11 January 2024).
- Boyer 2008** Boyer, A.-M. (2008) *Les paralittératures*. Paris: Armand Colin.
- Brecht 2004** Brecht, B. (2015) “Über die Popularität des Kriminalromans”, in J. Vogt (ed.) *Facetten des Kriminalromans*, pp. 315–321.
- Brezina 2018** Brezina, V. (2018) *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.
- Brezina and Meyerhoff 2014** Brezina, V. and Meyerhoff, M. (2014) “Significant or random?: A critical review of sociolinguistic generalisations based on large corpora”, *International Journal of Corpus Linguistics*, 19(1), pp. 1–28. Available at: <https://doi.org/10.1075/ijcl.19.1.01bre>.
- Burrows 2007** Burrows, J. (2007) “All the way through: Testing for authorship in different frequency strata”, *Literary and Linguistic Computing*, 22(1), pp. 27–47. Available at: <https://doi.org/10.1093/lcl/fqi067> (Accessed: 6 November 2024).
- Calvo Tello 2021** Calvo Tello, J. (2021) *The novel in the Spanish Silver Age: A digital analysis of genre using machine learning*. Bielefeld: transcript (Digital humanities research, volume 4).
- Chang 2009** Chang, J. et al. (2009) “Reading tea leaves: How humans interpret topic models”, in *Advances in neural information processing systems*, pp. 288–296.
- Constans 1999** Constans, E. (1999) *Parlez-moi d’amour: le roman sentimental : des romans grecs aux collections de l’an 2000*. Limoges: Presse Universitaire de Limoges.
- Couégnas 1992** Couégnas, D. (1992) *Introduction à la paralittérature*. Paris: Editions du Seuil (Collection Poétique).
- Craig and Kinney 2009** Craig, H. and Kinney, A.F. (eds). (2009) *Shakespeare, Computers, and the Mystery of Authorship*. 1st edn. Cambridge University Press.
- Culpeper 2009** Culpeper, J. (2009) “Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare’s Romeo and Juliet”, *International Journal of Corpus Linguistics*, 14(1), pp. 29–59. Available at: <https://doi.org/10.1075/ijcl.14.1.03cul> (Accessed: 6 November 2024).
- Damian-Gaillard 2011** Damian-Gaillard, B. (2011) “Les romans sentimentaux des collections Harlequin : quelle(s) figure(s) de l’amoureux? Quel(s) modèle(s) de relation(s) amoureuse(s)?”, *Questions de Communication*, (20), pp. 317–336. Available at: <https://doi.org/10.4000/questionsdecommunication.2130> (Accessed: 6 November 2024).
- Du and Schöch 2024** Du, K., & Schöch, C. (2024). *600 French Novels in Derived Text Format* (Version v1.0.0) [Data Set]. <https://github.com/Zeta-and-Company/df600> (Accessed: 6 November 2024).
- Du et al. 2021** Du, K. et al. (2021) “Zeta & Eta: An Exploration and Evaluation of Two Dispersion-based Measures of Distinctiveness”, *Proceedings Computational Humanities Research 2021*, 1613, p. 0073. Available at: [http://ceur-ws.org/Vol-2989/short\\_paper11.pdf](http://ceur-ws.org/Vol-2989/short_paper11.pdf).

- Du et al. 2022a** Du, K. et al. (2022) “Kontrastive Textanalyse mit pydistinto - Ein Python-Paket zur Nutzung unterschiedlicher Distinktivitätsmaße”, *Zenodo*, 7 March. Available at: <https://doi.org/10.5281/zenodo.6327967> (Accessed: 6 November 2024).
- Du et al. 2022b** Du, K., et al. (2022) “Evaluation of measures of distinctiveness: Classification of literary texts on the basis of distinctive words”, *Journal of Computational Literary Studies*, 1(1). Available at: <https://doi.org/10.48694/JCLS.102> (Accessed: 6 November 2024).
- Dubois 2005** Dubois, J. (2005) *Le roman policier ou la modernité*. Paris: Armand Colin.
- Dudovitz\_2019** Dudovitz, R.L. (2019) “The boundary between the romance and the bestseller: Harlequins, historical novels, and family sagas”, in *The Myth of Superwoman: Women’s Bestsellers in France and the United States*. Milton Park: Routledge.
- Dunning 1993** Dunning, T. (1993) “Accurate methods for the statistics of surprise and coincidence”, *Computational Linguistics*, 19(1), p. 14. Available at: <http://aclweb.org/anthology/J93-1003> (Accessed: 6 November 2024).
- Eder et al. 2016** Eder M., et al. (2016). “Stylometry with R: A package for computational text analysis”, *R Journal*, 8(1), 107–121. <https://journal.r-project.org/archive/2016/RJ-2016-007/index.html>.
- Franzini et al. 2018** Franzini, G. et al. (2018) “Attributing Authorship in the noisy digitized correspondence of Jacob and Wilhelm Grimm”, *Frontiers in Digital Humanities*, 5. Available at: <https://doi.org/10.3389/fdigh.2018.00004> (Accessed: 6 November 2024).
- Gittel and Köppe 2022** Gittel, B. and Köppe, T. (2022) “On the Distance Between Traditional and DH-Based Genre Theory”, *Digital Journal for Philology*. Available at: <https://doi.org/10.17879/64059431694>.
- Gonon et al. 2017** Gonon, L. et al. (2017) “Les phraséologismes spécifiques à deux sous-genres de la paralittérature: le roman sentimental et le roman policier” in *Colloque Phraséologie française*, Arras: Université d’Artois.
- Gonon et al. 2018** Gonon, L. et al. (2018) “Motifs textuels spécifiques au genre policier et à la littérature blanche”, *SHS Web of Conferences*. Edited by F. Neveu et al., 46. Available at: <https://doi.org/10.1051/shsconf/20184606007> (Accessed: 6 November 2024).
- Grimm and Hartwig 2016** Grimm, J. and Hartwig, S. (2016) *Französische Literaturgeschichte*. Berlin: Springer-Verlag.
- Grossman et al. 2020** Grossmann, F., et al. (2020) “Alcohol and tobacco consumption in English and French novels since the 1950s: A corpus-stylistic analysis”, in I. Novakova and D. Siepmann (eds.) *Phraseology and Style in Subgenres of the Novel: A Synthesis of Corpus and Literary Perspectives*. Cham: Springer International Publishing, pp. 115–150. Available at: [https://doi.org/10.1007/978-3-030-23744-8\\_5](https://doi.org/10.1007/978-3-030-23744-8_5) (Accessed: 6 November 2024).
- Heiden 2010** Heiden, S. (2010). The TXM Platform: Building open-source textual analysis software compatible with the TEI Encoding Scheme. In *24th Pacific Asia Conference on Language, Information and Computation* (pp. 389–398). Sendai, Japan.
- Helgorsky 1985** Helgorsky, F. (1985) “Harlequin ou la quête du grand amour”, *Communication et langages*, 63(1), pp. 83–98. Available at: <https://doi.org/10.3406/colan.1985.1666> (Accessed: 6 November 2024).
- Helgorsky 1987** Helgorsky, F. (1987) “Harlequin: l’unité dans la diversité et vice-versa”, *Pratiques*, 54(1), pp. 5–19. Available at: <https://doi.org/10.3406/prati.1987.1434> (Accessed: 6 November 2024).
- Hoover 2010** Hoover, D.L. (2010) “Teasing out authorship and style with t-tests and Zeta”, in *Digital Humanities Conference. Digital Humanities Conference*, London: ADHO.
- Hoover 2020** Hoover, D. L. (2020). “Authorial style”, in Dan McIntyre (ed.) *Language and Style*, pp. 250–271. London: Bloomsbury Publishing.
- Hoppenstand 2016** Hoppenstand, G. (2016) “Genres and formulas in popular literature”. In *A Companion to Popular Culture*. John Wiley & Sons, Ltd, 2016. <https://doi.org/10.1002/9781118883341.ch7>.
- Jensen 1980** Jensen, M.A. (1980) *Women and romantic fiction: A Case study of Harlequin enterprises, romances, and readers*. Thesis. Hamilton: McMaster University. Available at: <https://macsphere.mcmaster.ca/handle/11375/15523> (Accessed: 6 November 2024).
- Kessler et al. 1997** Kessler, B., et al. (1997) “Automatic detection of text genre”, in *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics. ACL-EACL 1997*, Madrid, Spain: Association for Computational Linguistics, pp. 32–38. Available at: <https://doi.org/10.3115/976909.979622> (Accessed: 6 November 2024).

- Kestemont 2018** Kestemont, M. et al. (2018) "Overview of the Author Identification Task at PAN-2018: Cross-domain Authorship Attribution and Style Change Detection" in. *Conference and Labs of the Evaluation Forum*. Aachen: CEUR-WS.
- Kilgarrriff 2001** Kilgarrriff, A. (2001) "Comparing Corpora", *International Journal of Corpus Linguistics*, 6(1), pp. 97–133. Available at: <https://doi.org/10.1075/ijcl.6.1.05kil> (Accessed: 6 November 2024).
- Kraif and Diwersy 2012** Kraif, O. and Diwersy, S. (2012) "Le Lexicoscope : un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques" in *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*. Volume 2. TALN, Grenoble: ATALA/AFCP, pp. 399–406. Available at: <https://paperswithcode.com/paper/le-lexicoscope-un-outil-pour-letude-de> (Accessed: 6 November 2024).
- Leclerc 1989** Leclerc, Y. (1989) "Autour de Minuit", *Dalhousie French Studies*, 17, pp. 63–74.
- Lijffijt 2014** Lijffijt, J. et al. (2014) "Significance testing of word frequencies in corpora", *Digital Scholarship in the Humanities*, 31(2), pp. 374–397. Available at: <https://doi.org/10.1093/llc/fqu064> (Accessed: 6 November 2024).
- Lits 1999** Lits, M. (1999) *Le roman policier: introduction à la théorie et à l'histoire d'un genre littéraire*. Liège: Editions du CEFAL.
- Masis 2023** Masis, Serg (2023) *Interpretable Machine Learning with Python*. Birmingham: Packt Publishing Ltd.
- Mather and Rheault 2016** Mather, P. and Rheault, S. (2016) *Rediscovering French Science-Fiction in Literature, Film and Comics: From Cyrano to Barbarella*. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Migozzi 2005** Migozzi, J. (2005) *Boulevards du populaire*. Limoges: Presse Universitaire de Limoges.
- Miller 2020** Miller, D. (2020) "Analysing Frequency Lists" in M. Paquot and S.Th. Gries (ed.) *A Practical Handbook of Corpus Linguistics*. Cham: Springer International Publishing, pp. 77–97. Available at: [https://doi.org/10.1007/978-3-030-46216-1\\_4](https://doi.org/10.1007/978-3-030-46216-1_4) (Accessed: 6 November 2024).
- Millet and Labbé 2001** Millet, G. and Labbé, D. (2001) *La science-fiction*. Paris: Belin (Sujets).
- Montani et al. 2023** Montani, I. et al. (2023) "explosion/spaCy: v3.7.2: Fixes for APIs and requirements". *Zenodo*. Available at: <https://doi.org/10.5281/zenodo.10009823> (Accessed: 6 November 2024).
- Nagy 2024** Nagy, Ben (2024). "Bootstrap distance imposters: High precision authorship verification with improved interpretability", in Wouter Haverals, Marijn Koolen and Laure Thompson (ed.) *Proceedings of the Computational Humanities Research Conference 2024*. December 4-6, 2024. Aarhus: CEUR-WS. Available at: <https://ceur-ws.org/Vol-3834/paper61.pdf> (Accessed: 11 January 2024).
- Niwattanakul et al. 2013** Niwattanakul, S., et al. (2013, March). "Using of Jaccard coefficient for keywords similarity", in *Proceedings of the International Multiconference of Engineers and Computer Scientists* (Vol. 1, No. 6, pp. 380-384).
- Olivier-Martin 2013** Olivier-Martin, Y. (2013) *Histoire du roman populaire en France de 1840 à 1980*. Paris: Albin Michel.
- Organisciak and Downie 2021** Organisciak, P. and Downie, J.S. (2021) "Research access to in-copyright texts in the humanities" in *Information and Knowledge Organisation in Digital Humanities*. Milton Park: Routledge. Available at: <https://doi.org/10.4324/9781003131816-8> (Accessed: 6 November 2024).
- Paquot and Bestgen 2009** Paquot, M. and Bestgen, Y. (2009) "Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction" in A.H. Jucker, D. Schreier, and M. Hundt (ed.) *Corpora: Pragmatics and Discourse*. Leiden: Brill | Rodopi. Available at: [https://doi.org/10.1163/9789042029101\\_014](https://doi.org/10.1163/9789042029101_014) (Accessed: 6 November 2024).
- Piper 2016** Piper, A. (2016) "Fictionality", *Journal of Cultural Analytics*. Available at: <https://doi.org/10.22148/16.011>.
- Rabine 1985** Rabine, L.W. (1985) "Romance in the age of electronics: Harlequin enterprises" in *Feminist Studies*, 11(1), pp. 39–60. Available at: <https://doi.org/10.2307/3180131> (Accessed: 6 November 2024).
- Rayson 2005** Rayson, Paul. (2005) *Wmatrix: A web-based corpus processing environment*. Lancaster, UK: Computing Department, Lancaster University.
- Rebora 2023** Rebora, S. (2023) "GPT-3 vs. Delta. Applying Stylometry to Large Language Models" in *La memoria digitale: forme del testo e organizzazione della conoscenza. Atti del XII Convegno Annuale AIUCD*, Siena: AIUCD, pp. 292–297.
- Ruxton 2006** Ruxton, G.D. (2006) "The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test", *Behavioral Ecology*, 17(4), pp. 688–690. Available at: <https://doi.org/10.1093/beheco/ark016>

(Accessed: 6 November 2024).

- Rybicki et al. 2014** Rybicki, J., et al. (2014) “Collaborative authorship: Conrad, Ford and Rolling Delta”, *Literary and Linguistic Computing*, 29(3), pp. 422–431. Available at: <https://doi.org/10.1093/lc/fqu016> (Accessed: 6 November 2024).
- Savoy 2020** Savoy, J. (2020) “Elena Ferrante: A case study in authorship attribution” in Savoy, J., *Machine Learning Methods for Stylometry*. Cham: Springer International Publishing, pp. 191–210. Available at: [https://doi.org/10.1007/978-3-030-53360-1\\_8](https://doi.org/10.1007/978-3-030-53360-1_8) (Accessed: 6 November 2024).
- Schöch et al. 2018** Schöch, C. et al. (2018) “Burrows' Zeta: Exploring and evaluating variants and parameters” in *Book of Abstracts of the Digital Humanities Conference. Digital Humanities Conference (DH2018)*, Mexico City: ADHO.
- Schöch 2018** Schöch, C. (2018) “Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie”, in T. Bernhart et al. (ed.) *Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven*. Berlin: de Gruyter, pp. 77–94.
- Schöch et al 2020** Schöch, C. et al. (2020) “Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen”, *Zeitschrift für digitale Geisteswissenschaften (ZfdG)*, 5. Available at: [http://dx.doi.org/10.17175/2020\\_006](http://dx.doi.org/10.17175/2020_006) (Accessed: 6 November 2024).
- Schröter et al. 2021** Schröter, J. et al. (2021) “From Keyness to Distinctiveness – Triangulation and Evaluation in Computational Literary Studies”, *Journal of Literary Theory*, 15(1–2), pp. 81–108. Available at: <https://doi.org/10.1515/jlt-2021-2011> (Accessed: 6 November 2024).
- Scott 1997** Scott, M. (1997) “PC Analysis of key words and key key words”, *System*, 25(2), pp. 233–245. Available at: [https://doi.org/10.1016/S0346-251X\(97\)00011-0](https://doi.org/10.1016/S0346-251X(97)00011-0) (Accessed: 6 November 2024).
- Scott 2009** Scott, M. (2009) “In search of a bad reference corpus”, in D. Archer (ed.) *What's in a Word-list? Investigating Word Frequency and Keyword Extraction*. Farnham, England ; Burlington, VT: Ashgate Pub. Ltd (Digital research in the arts and humanities), pp. 80–91.
- Scott 2024** Scott, M., 2024, *WordSmith Tools version 9 (64 bit version) Stroud: Lexical Analysis Software*. Available at: <https://lexically.net/wordsmith/> (last accessed 27 August 2025).
- Slusser 1989** Slusser, G. (1989) “‘Science Fiction in France’: An introduction”, *Science Fiction Studies*, 16(3), pp. 251–253.
- Thomas 1989** Thomas, P.J. (1989) “The Current state of science fiction in France (L'état actuel de la science-fiction en France)”, *Science Fiction Studies*, 16(3), pp. 298–306.
- Todorov 1971** Todorov, C. (1971) *Poétique de la prose / Tzvetan Todorov*. Paris: Ed. du Seuil (Collection poétique).
- Tranchese 2023** Tranchese, A. (2023) *From Fritzl to #metoo: Twelve years of Rape Coverage in the British Press*. Berlin: Springer Nature.
- Underwood 2019** Underwood, T. (2019) *Distant horizons: digital evidence and literary change*. Chicago: The University of Chicago Press.
- Vanoncini 2002** Vanoncini, A. (2002) *Le roman policier*. 3. éd., mise à jour. Paris: Presses Univ. de France (Que sais-je ?, 1623).
- Welch 1947** Welch, B.L. (1947) “The generalization of Student's problem when several different population variances are involved”, *Biometrika*, 34(1–2), pp. 28–35. Available at: <https://doi.org/10.1093/biomet/34.1-2.28> (Accessed: 6 November 2024).
- Worsham and Kalita 2018** Worsham, J. and Kalita, J. (2018) “Genre identification and the compositional effect of genre in literature”, in *Proceedings of the 27th International Conference on Computational Linguistics. COLING 2018*, Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 1963–1973. Available at: <https://aclanthology.org/C18-1167> (Accessed: 6 November 2024).
- Yule 1944** Yule, C.U. (1944) *The statistical study of literary vocabulary*. Cambridge: Cambridge University Press.

## Recommendations

DHQ is testing out three new article recommendation methods! Please explore the links below to find articles that are related in different ways to the one you just read. We are interested in how these methods work for readers—if you would like to share feedback with us, please complete our short evaluation survey. You can also visit our documentation

for these recommendation methods to learn more.

## SPECTER Recommendations

Below are article recommendations generated by the SPECTER model:

1. J. M. Coetzee's Work in Stylostatistics, 2014, Peter Johnston, Royal Holloway, University of London
2. Interpreting Measures of Meaning: Introducing Salience Differentiated Stability, 2022, Hugo Dirk Hogenbirk, University of Groningen, Departement of the History of Philosophy; Wim Mol
3. Computational Linguistics and Classical Lexicography, 2009, David Bamman, Tufts University; Gregory Crane, Tufts University
4. Vive la Différence! Text Mining Gender Difference in French Literature, 2009, Shlomo Argamon, Linguistic Cognition Lab, Dept. of Computer Science, Illinois Institute of Technology; Jean-Baptiste Goulain, Linguistic Cognition Lab, Dept. of Computer Science, Illinois Institute of Technology; Russell Horton, Digital Library Development Center, University of Chicago; Mark Olsen, ARTFL Project, University of Chicago
5. In a test bed with Kafka. Introducing a mixed-method approach to digital stylistics, 2018, J. Berenike Herrmann, Basel University

## DHQ Keyword Recommendations

Below are article recommendations generated by DHQ Keywords:

1. Exploring Combinatorial Methods to Produce Sonnets: An Overview of the Oupoco Project, 2024, Frédérique Mélanie-Becquet, LATTICE (CNRS & ENS/PSL & Univ. Sorbonne Nouvelle); Clément Plancq, LATTICE (CNRS & ENS/PSL & Univ. Sorbonne Nouvelle); Claude Grunspan, LATTICE (CNRS & ENS/PSL & Univ. Sorbonne Nouvelle); Mylène Maignant, LATTICE (CNRS & ENS/PSL & Univ. Sorbonne Nouvelle); Matthieu Raffard, Atelier Raffard-Roussel; Mathilde Roussel, Atelier Raffard-Roussel; Fiammetta Ghedini, RIVA Illustrations; Thierry Poibeau, LATTICE (CNRS & ENS/PSL & Univ. Sorbonne Nouvelle)
2. Reading the Quan Tang shi: Literary History, Topic Modeling, Divergence Measures, 2019, Peter Broadwell, Center for Interdisciplinary Research, Stanford University; Jack W. Chen, University of Virginia; David Shepard, Scholarly Innovation Lab, UCLA
3. GIS and Literary History: Advancing Digital Humanities research through the Spatial Analysis of historical travel writing and topographical literature, 2016, Patricia Murrieta-Flores, Digital Humanities Research Centre, University of Chester; Christopher Donaldson, History Department, Lancaster University; Ian Gregory, History Department, Lancaster University
4. Discourse cohesion in Xenophon's On Horsemanship through Sketch Engine, 2023, Victoria Beatrix Fendel, University of Oxford; Matthew T. Ireland, Sidney Sussex College, University of Cambridge
5. Sentiment Analysis: Limits and Progress of the Syuzhet Package and Its Lexicons, 2022, Hyeol Kim, Texas A&M University

## TF-IDF Recommendations

Below are article recommendations generated by the TF-IDF Model:

1. Mining for the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers, 2014, Carolyn Strange, Australian National University; Daniel McNamara; Josh Wodak, Australian National University; Ian Wood, Research School of Computer Science, Australian National University
2. The Stylometry of Film Dialogue: Pros and Pitfalls, 2020, Agata Hołobut, Jagiellonian University in Kraków; Jan Rybicki, Jagiellonian University in Kraków
3. As You Can See: Applying Visual Collaborative Filtering to Works of Art, 2008, Gerhard Jan Nauta, Leiden University
4. Readies Online, 2011, Craig Saper, University of Maryland, Baltimore County
5. Revenge of the Nerd: Junot Díaz and the Networks of American Literary Imagination, 2013, Ed Finn, Arizona State University



This work is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International License.