Stylometric similarity in literary corpora: Non-authorship clustering and Deutscher Novellenschatz

Journal Article

Author(s):
Papcke, Simon; Weitin, Thomas; Herget, Katharina; Glawion, Anastasia; Brandes, Ulrik

Publication date:
2023-04

Permanent link:
https://doi.org/10.3929/ethz-b-000564866

Rights / license:
Creative Commons Attribution 4.0 International

Originally published in:
Digital Scholarship in the Humanities 38(1), https://doi.org/10.1093/llc/fqac039

Stylometric similarity in literary corpora: Non-authorship clustering and *Deutscher Novellenschatz*

Simon Päpcke
Social Networks Lab, ETH Zurich, Zurich, Switzerland

Thomas Weitin, Katharina Herget and Anastasia Glawion
LitLab, TU Darmstadt, Darmstadt, Germany

Ulrik Brandes
Social Networks Lab, ETH Zurich, Zurich, Switzerland

Correspondence: Simon Päpcke, Social Networks Lab, ETH Zurich, Zurich, Switzerland.
E-mail:
simon.paepcke@gess.ethz.ch

Abstract

A distant-reading task in literary corpus analysis is to group stylometrically similar texts. Since there are many ways to define writing style, the result not only depends on the clustering method but even more so on the measure of similarity. With authorship attribution, the predominant application of stylometry, as its benchmark much research has addressed the utility of methods for measuring similarity. We use a corpus of German-language novellas to demonstrate that one may be interested in very different meaningful groups of texts simultaneously, and that these can be recovered from stylometric clustering if the measure is chosen accordingly. As can be expected, different measures do better at recovering groups associated with, for instance, subgenre, author gender, or narrative perspective. As a consequence, it is suggested that corpus analyses should not be based on what is currently considered the most refined measure of stylometric similarity, but rather break down the decisions that yield a specific measure and provide substantively justified arguments for them.

1 Introduction

Authorship attribution from stylometric similarity rests on the assumption that there is an immutable signal that authors emit involuntarily. This signal is often claimed to manifest itself in the use of function words. The utility derived from an author invariant is that it yields a higher similarity between texts from the same author than between texts from different authors, so that authorship can be recovered from clustering similar texts.

Since stylometric similarity can be measured in different ways it appears a fair question to ask which measure is most suitable? Indeed, scholars have championed various measures with strong support articulated, for example, for Burrows's Delta (Burrows, 2002) and Zeta (Burrows, 2007; Schöch *et al.*, 2018). The arguments are largely derived from empirical findings about relative performance, however, and therefore do not necessarily generalize.

Moreover, the same measures have also been used for stylometric analyses of corpora not aimed at

https://doi.org/10.1093/llc/fqac039 Advance Access published on 9 August 2022

authorship but, for instance, genre discrimination (Schöch, 2014). Although individual analyses led to convincing results, it is implausible that signals underlying different categorizations should be comparable and that one notion of stylometric similarity should serve multiple classification purposes.

The purpose of this article is therefore two-fold. First, we want to point out decision points in the selection of similarity measures. Secondly, we want to demonstrate that, indeed, specifically designed similarity measures are capable of identifying a variety of meaningful groups of texts in a literary corpus, beyond authorship. To keep the discussion concentrated, we restrict ourselves to stylometric similarity in terms of word occurrences and focus on a single corpus, the *Deutscher Novellenschatz*, published 1871–1876 and edited by Heyse and Kurz. The corpus is particularly suited for our purpose as it is not a representative sample of novellas but the result of a historical process during which the editors aimed for a canonical collection and were acutely aware of compositional effects. Using different stylometric similarity measures, recent studies already found the two novellas of the editors to be central in different clusters of a similarity network (Weitin, 2016), and a group of novellas that appear to have been influenced stylistically by Eichendorff's very last novella, *Die Glücksritter* (Jannidis, 2017).

The remainder of the article is structured into three parts. We start with some additional background on our focal corpus. The main part breaks down the decisions made during the construction of a similarity measure, interspersed with smaller scale examples taken from the *Novellenschatz*. In the final part, we explore groupings of novellas obtained from different similarity measures. We conclude with a discussion of potential consequences for the quantitative study of literary corpora.

2 Background on the Corpus

The corpus of *Deutscher Novellenschatz* consists of twenty-four volumes published between 1871 and 1876. It contains $N = 86$ novellas that have originally been published between 1811 and 1875 in a variety of literary contexts. Its editors, Paul Heyse and Hermann Kurz, wanted the *Novellenschatz* to be a paradigmatic sample of the novella style and have therefore forgone any original contributions. Their selection became a bestseller despite the fact that it did not contain new material and was rather expensive. Heyse followed up almost immediately with two more Novellenschatz collections with the same publisher, and editors of other collections tried to copy the success.

In fact, the term 'Novellenschatz' became an epitome of the genre's proliferation, and when the century had turned and realism itself had become literary history, belonging to such a collection proved sufficient to identify a literary text as a novella. In the introduction of the *Deutscher Novellenschatz*, however, the genre is characterized as having a simple plot with a reduced character set and an easily recognizable symbolism.

Given the historical poetics of the genre, Jannidis (2017) compares the novellas of the *Novellenschatz* to a corpus of novels. Construction networks in which nodes represent characters and edges represent co-references to them, the index of degree centrality is determined to assess the importance of characters. Findings support the idea that because of the restricted character set within a novella very few main figures 'absorb' all centrality whereas for the novel with its many figures centrality concentration is lower.

Weitin and Herget (2017) use topic models to explore deep semantic structures in the corpus and compared them with evidence from close reading. On the one hand, topics about religion, the justice system, economic issues, and rural life consolidate what is known of the novella of German-language realism. On the other hand, the subject of marriage that shapes the plot of a vast majority of the novellas of the epoch does not appear as a topic so much. A number of topics consisted exclusively of words characteristic for single novellas even when the analyses uses chunked versions of the texts. To label this effect, which no parameter manipulation could eliminate, and following Heyse's theoretical postulate that novellas must be summarizable in a few words, the term 'falcon topics' is coined.

Weitin (2016) uses a network model for the entire corpus in which each novella is a node, and edges represent differences in the use of frequent words. The edge attribute is defined as Burrows's Delta between the two texts, and a node attribute is created assessing the distance between the text and a corpus average. Heyse's novella *Der Weinhüter von Meran* has the lowest distance from the corpus average and is central for a large group of novellas, thus embodying a stylometric average of the *Novellenschatz*. In contrast, Kurz's very last novella *Die beiden Tubus* (which Heyse gave a final form after the sudden death of his co-editor) is not average at all but takes a central position within a subset of novellas. Weitin (in press) extended the analysis by identifying the words contributing the most to the observed stylometric differences. One finding is that inflexions of the pronoun 'I' have high standard deviation and can indeed be used to detect first-person narratives. Furthermore, a graph of the mean delta distance of the corpus texts and their entropy can help to understand not only the degree but also the quality of stylometric differences and similarities.

3 Stylometric Similarity

Stylometry uses quantification to study writing style. Here, we focus specifically on the frequencies of common words as a means to group the novellas in the *Novellenschatz* by similarity.

Document similarity in terms of common words is an established concept in information retrieval (Baeza-Yates and Ribeiro-Neto, 2011). It is often considered in the abstract, with evaluation performed on generic document collections. Our interest, however, is not in the discovery of a specific group of document clusters but in demonstrating differences among results obtained by different methods.

It is worth noting that authorship attribution is a distinct problem for which it is indeed meaningful to compare methods in terms of their ability to identify texts of the same authors with high accuracy. There is less reason to assume that a method designed specifically to discriminate between authors would also serve well to identify stylometric differences across epochs, genres, and other grouping criteria. In fact, this would not even be desirable because of interaction effects between author invariants and other signals.

Instead of proposing a particular notion of (dis)-similarity, we therefore break down the process of determining similarity into a number of generic steps. Our hope is to thus inform the construction of similarity measures tailored to specific interests.

In the following, we denote with $\mathcal{C} = \{D_1, \ldots, D_N\}$ a collection (the corpus) of documents (the texts). To determine pairwise similarities, each document is characterized by an n-dimensional feature vector $t(D) \in \mathbb{R}^n$ that will be derived from the occurrences of words in document D. A distance $\delta : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ in the corresponding vector space defines then their dissimilarity.

3.1 Features

To keep the discussion focused, we characterize texts only by features that originate from word frequencies. The large set of stylometric features (Stamatatos, 2009) thus excluded from our discussion includes n-grams, co-occurrences, word or sentence length distributions, and sentence complexity.

Even in this restricted setting, a number of choices have to be made. Unlike suggested by the quest for ever more accurate authorship attribution, their relative merit may change with context. A choice may lead to superior analysis in one respect but fail to do so in another. It is thus worthwhile to consider multiple options and evaluate their consequences.

3.1.1 *Canonicalization*

The first major decision is the granularity at which lexical items are distinguished and whether different meanings of the related items are considered. The resulting classes of lexical items that are treated as equivalent constitute our feature variables $t_1, \ldots, t_n$. We refer to them as *terms*, even when they represent entire classes of character strings because these are treated as equivalent manifestations of the same unit of observation.

For instance, we may use all-lowercase to avoid distinguishing occurrences of words at the beginning or in the middle of a sentence. On the other hand, we may still want to retain the distinction between the capitalized version of the German word 'liebe' (dear, lovely) and the noun 'Liebe' (love).

> Gnädiger Herr, antwortete die Frau mit neuer Betrübniß, meine Liebe trägt die Schuld von alle dem Unglück [. . .]
>
> Arnim, *Der tolle Invalide auf dem Fort Ratonneau*

> Warum wollen Sie so rasch fort von hier, liebe Emma?
>
> Grimm, *Das Kind*

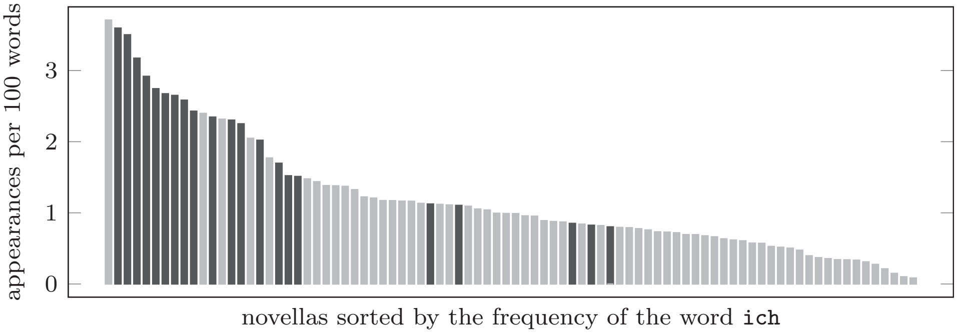


Fig. 1 Relative frequencies of the word `ich` in the novellas with first-person narratives marked in black. Without changing pronouns to a generic term, we expect a strong connection of first-person narratives as an artifact of the method.

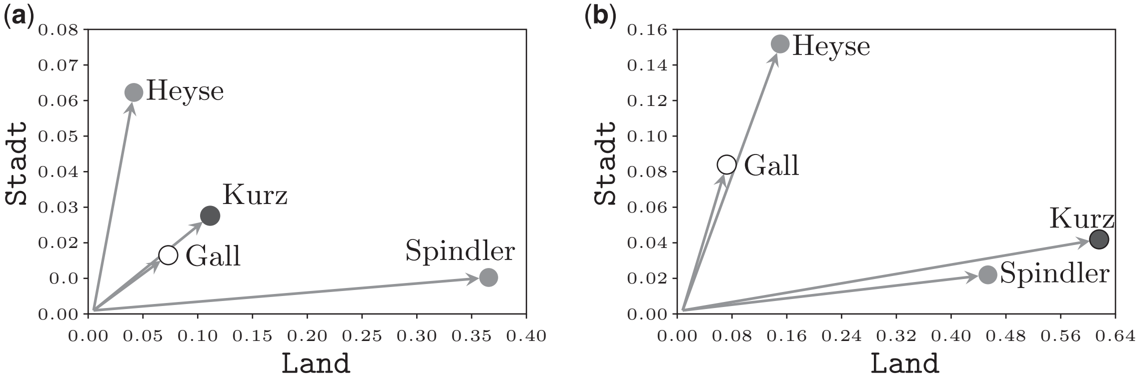


Fig. 2 By using a dictionary to map named locations to classes identifiers of locations we note a change of document frequency vectors of Kurz's *Die beiden Tubus* and Gall's *Eine fromme Lüge*. Up to scaling, there is a strong similarity between the vectors of Kurz and Gall on the left. However, merging particular places into generic classes leads to a situation in which the novellas of Gall and Kurz are rather dissimilar and much closer to the texts of Heyse and Spindler, respectively. (a) term frequency vectors for the words `Stadt` (urban setting) and `Land` (rural setting) in four novellas and (b) the same vectors after replacing named locations by the appropriate generic term `Stadt` or `Land`.

Similarly, stemming, lemmatization, and even more inclusive abstractions are often used to eliminate undesired distinctions (Fig. 1).

Multiple lexically different references to a named entity such as 'Meran' can be counted toward that same entity, or not. As a consequence, the number of occurrences of a particular place (Meran), of a class of places (city), or of any kind of place (locations) yield different features and thus lead to different similarities later on (Fig. 2).

The degree to which syntactic, semantic, and contextual information is used to aggregate or disaggregate the words of a document into classes that define term features t_i, $i = 1, \ldots, n$, will have a profound impact on which texts are found to be similar.

3.1.2 *Counting occurrences*

With the units of observation decided upon, we turn to determining their associated values. A straightforward measurement considers raw counts: each time a member of an equivalence class (say, all words associated with the infinitive 'to be') is encountered, it contributes one unit toward the value of the associated variable.

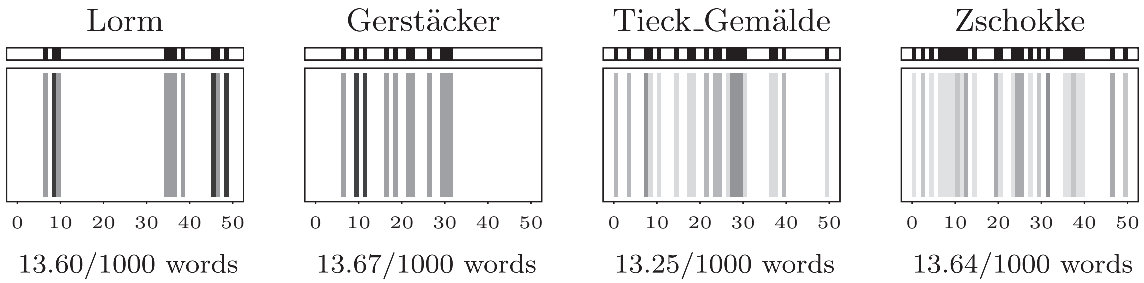


Fig. 3 Appearance profile of the word `Vater` (father). Each novella is split into fifty equal slices and we denote a gray bar if the term appears in a slice and a white bar otherwise where darker color indicates a higher accumulated appearance in the slice. In Lorm's *Ein adeliges Fräulein* and Gerstäcker's *Germelshausen*, we can identify the present story within a story by the accumulated occurrence of the term that is not present in the texts of Tieck and Zschokke.

Raw counts do not discriminate between occurrences of a particular word that are highly concentrated in one section, or spread over the entire text. In variations, a text may therefore be chunked into segments, say of equal length, and each segment is quantified separately, or one occurrence is counted if and only if the raw count or relative frequency in a segment surpasses a threshold (Fig. 3).

While the later analysis is highlighting stylometric and content-wise differences within a text, the former is suited for highlighting stylometric differences between texts. To keep the discussion focused, we will concentrate on the differences between texts. In the following, we will use $t_i(D_k)$ to denote the count of the ith term (equivalence class of words) obtained for the kth document (corpus text).

In this example, we have mapped `Basel`, `Wien`, `Bordeaux`, `Laibach`, `Verona`, `Meran`, `Innsbruck`, `Venedig`, `Berlin`, and `Bremen` to `Stadt`, whereas `A...berg`, `Appenzell`, `Aarlberg`, `Bernerland`, `Y...burg`, `Sch...ingen`, `Tirol`, `Burgland`, `Etschtal`, `Küchelberg`, `Vitschgau`, `Algrund`, and `Trautmannsdorf` to `Land`.

3.1.3 Filtering

For common levels of granularity the number of terms in a typical document collection is rather large. It usually includes many elements that occur only rarely or are not informative for other reasons. Function words, for instance, are considered essential elements in stylometric analysis for authorship attribution but irrelevant in topic analysis. Terms that appear only occasionally or in few documents may be informative or constitute noise. Weitin (in press) used entropy analysis to illustrate this point.

Filtering serves to obtain a dimensionality, n, of the feature vector that is much smaller than the total number of terms actually present. It is often based on a dictionary (e.g. blacklisting of stop words) or numbers of occurrences (e.g. top-n most frequent words) (Fig. 4).

Culling is a variant of the frequency-based approach in which a word must, in addition to being frequent, appear in a minimum number of documents (Fig. 5).

3.2 Normalization

After determining features by deciding what to count, the comparison of texts by their feature vectors requires adjustments in order to take into account that frequencies may have different baselines in different texts.

A straightforward quantity to control for is the length of the text from which a feature is derived. A common normalization is therefore the share of occurrences counted toward a particular feature,

$$\mathrm{tf}(i, D) = \frac{t_i(D)}{|D|} ,$$

where $|D|$ is the text length of a document D in the corpus. The resulting quantity is called 'relative term frequency'. Although length-normalization yields a distribution of occurrences, we may need to establish a corpus baseline to identify the special role of a text in a corpus. In information retrieval, this is often done by multiplying relative frequencies with the

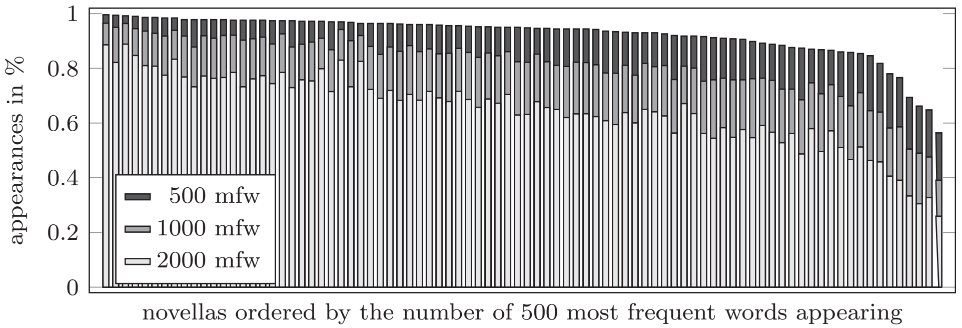
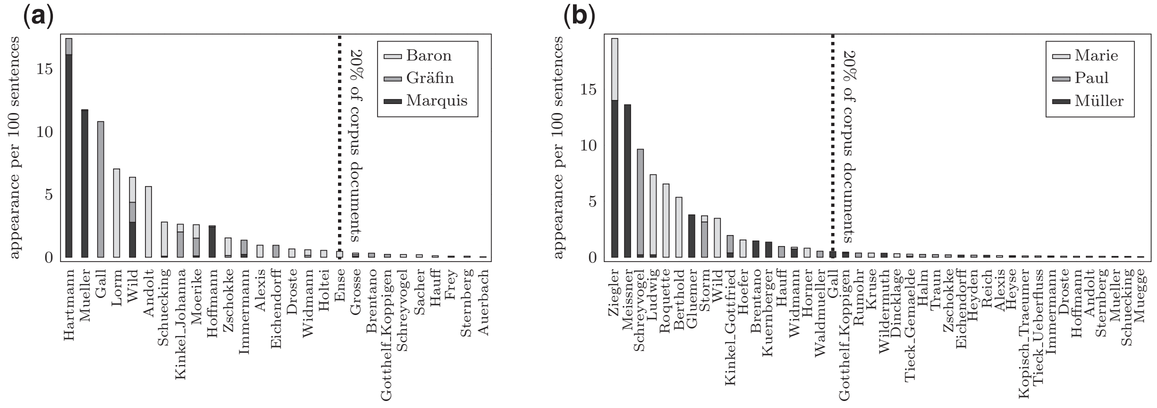


Fig. 4 While nearly every word in the 100 most frequent words occurs in every text, differences are more pronounced in the tail of the most frequent word vector. Novellas containing many words in the tail are affected by an increased dimensionality.



`Baron`, `Gräfin` and `Marquis`

`Marie`, `Müller` and `Paul`

Fig. 5 Names make up twenty-seven of the thirty-four words removed by 20% culling. Three of the other seven are `Baron`, `Gräfin`, and `Marquis`. These titles are present each in fifteen or sixteen of the eighty-six novellas and therefore close to the cutoff, which they pass if combined. A similar situation arises for the three most common names. While there may be substantive reasons such as a genre signal to combine noble titles, this is not expected to be equally meaningful for character names. (a) `Baron`, `Gräfin`, and `Marquis`; (b) `Marie`, `Müller`, and `Paul`.

logarithmically scaled inverse of the share of documents in which a term appears,

$$\mathrm{tf}-\mathrm{idf}(i, D) = \mathrm{tf}(i, D) \cdot \log \frac{N}{N(i)} \ ,$$

where $N = |\mathcal{C}|$ denotes the number of documents in the corpus and $N(i) = |\{k = 1, \dots, N : t_i(D_k) > 0\}|$ the number of documents in which the term i appears.[1] This weighting, and minor variants thereof, are referred to as 'inverse document frequency' (Manning *et al.*, 2008). It serves to boost terms that appear in fewer documents and thus are potentially more informative to discriminate documents (see also Fig. 6). Note that the weight of terms appearing in all documents equals zero and that the weighting by inverse document frequency is constant, and thus irrelevant, when all n terms appear in equal numbers of documents.

A different weighting is obtained by comparing relative frequencies across documents. Assuming a normal distribution we let $\mu(\mathrm{tf}(i))$ denote the expected relative frequencies of term i across all documents, i.e. the average over the entire corpus, and $\sigma(\mathrm{tf}(i))$ its standard deviation. Then,

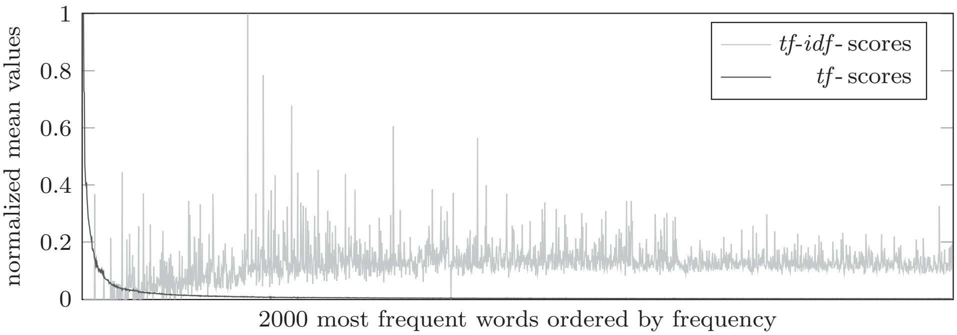


Fig. 6 We plot the mean *tf*- and *tf-idf*-score over all novellas. We observe the well-established Zipf's law (Zipf, 1935, 1949) for the *tf*-scores. Hence, any analysis that uses absolute differences of this score will be highly skewed toward the very frequent words while the *tf-idf*-score is reaching its peak at the medium frequent words. Meanwhile, the *z*-score does not discriminate between frequent and unfrequent words. In fact, its mean is zero by definition.

$$z_i(D) = \frac{\mathrm{tf}(i, D) - \mu(\mathrm{tf}(i))}{\sigma(\mathrm{tf}(i))}$$

defines the *z*-score of the *i*th term in document *D*. It is thus positive (negative), if the relative frequency of $\mathrm{tf}(i, D)$ in *D* is higher (lower) than expected across the corpus, where differences are scaled by their standard deviation, and thus made comparable. Since *z*-scores vary greatly for rare terms, they are generally used for the most frequent terms only. See Fig. 7 for a comparison.

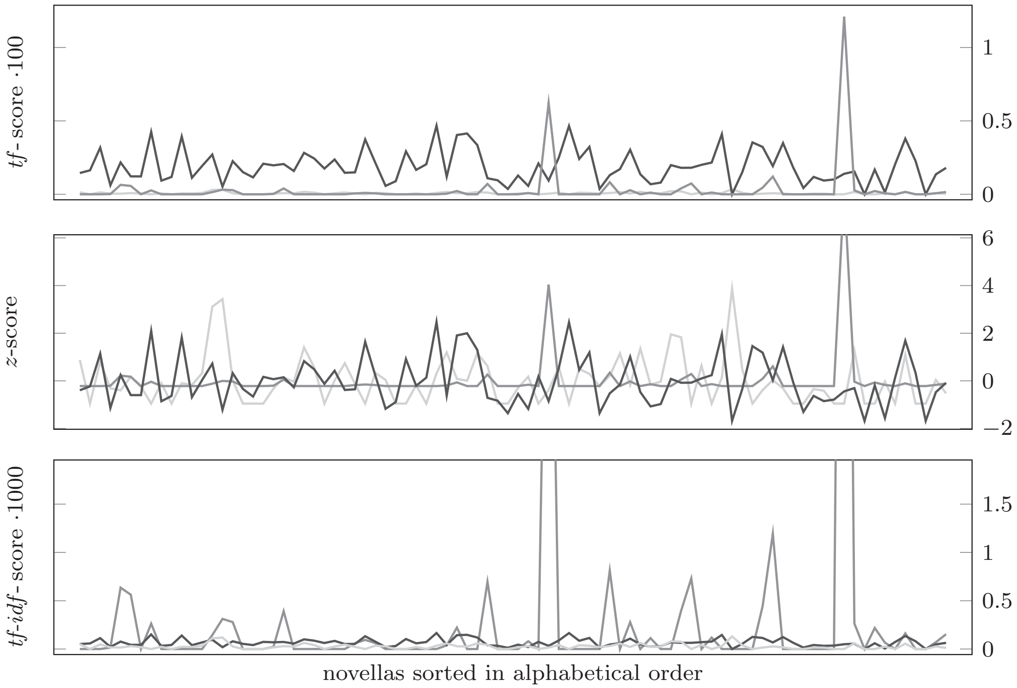
It may seem so far that normalization is largely with respect to a document or the corpus. For relative frequencies of all terms in any document $D \in \mathcal{C}$ we have, of course, $\sum_i \mathrm{tf}(i, D) = \sum_i t_i(D)/|D| = 1$, i.e. the relative frequencies of all terms sum to one in each document. Depending on the distances used later on, normalization may be more suitable if with respect to a different norm. With Euclidean distance, for instance, we obtain a vector of unit length for *D* from the division of each entry by its vector length $||t(D)||_2 = \sqrt{\sum_{i=1}^{n} t_i(D)^2}$. In comparison to relative frequencies, the use of Euclidean unit-length normalizations puts relatively more weight on larger deviations. In addition to the above within-vector normalizations, Büttner *et al.* (2017) list two variants that use thresholding. The first introduces a lower and an upper bound and clips frequencies outside of this interval to its boundaries. In a stronger discretization, the second variant replaces each entry by −1, 0, or 1 depending on whether the term is infrequent, standard, or frequent relative to the document or the entire corpus.

3.3 Dissimilarity

The basic question for which we are trying to find quantitative answers is whether two texts are similar with respect to the prevalence of words. We have already discussed that such comparison requires us to be precise about the term features that words are aggregated into, and the way we normalize frequencies across a document, the corpus, or with respect to each other. However, we also have to be clear about how to compare the frequencies of each term and how to aggregate their individual differences. It is a task-specific question, for instance, whether large differences with respect to some specific terms imply the same level of dissimilarity as many small differences across the board.

Options to assess such trade-offs have been discussed extensively (Burrows, 2002; Argamon, 2008; Smith and Aldridge, 2011; Sidorov *et al.*, 2014) although generally in attempts to establish the superiority of one measure over others. We briefly review the most commonly used concepts.

The distance of two numbers $a, b \in \mathbb{R}$ is computed as $|a - b|$. In mathematics, a norm is an extension of the absolute value that can, among other purposes, be used to translate the concept of a distance to multidimensional objects. In stylometric analysis we can measure the distance of two feature vectors $x, y \in$



	μ (σ)	*tf*- score ·100	*z*-score	*tf-idf*- score ·1000
—	`ihre`	0.1891 (0.1126)	0.0 (1.0)	0.0672 (0.0400)
—	`Pfarrer`	0.0317 (0.1460)	0.0 (1.0)	0.3139 (1.4434)
—	`eigenes`	0.0069 (0.0072)	0.0 (1.0)	0.0259 (0.0271)

Fig. 7 Comparison of the *tf-idf*-score, the *tf* -score and the *z*-score of `ihre` (63rd mfw), `Pfarrer` (378th mfw), and `eigenes` (1247th mfw). While the *tf-idf*-score is low for the frequent and the unfrequent words `ihre` and `eigenes` it is high for the medium frequent word `Pfarrer` which is very distinct for certain texts. In comparison, the term frequency of `ihre` is much higher for almost all novellas than for the other terms. With the *z*-score, these differences are balanced out by design.

$\mathbb{R}^n$ that result from some selection of terms and a normalization of their occurrence counts (e.g. *tf*-scores or *z*-scores) by

$$\delta_p(x, y) = \frac{1}{n} \cdot ||x - y||_p^p = \frac{1}{n} \sum_{i=1}^{n} |x_i - y_i|^p \ ,$$

where $||x - y||_p = (\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p}$ is the so-called p-norm for $p \geq 1$. For $p = 1$, we obtain the average absolute difference over all feature values also known as the Manhattan distance. Burrows's Delta (Burrows, 2002) is the application of δ_1 to feature vectors that consist of the z-scores for the n most frequent words.

For $p = 2$ we get the distance in Euclidean space, and δ_2 is also referred to as Quadratic Delta in the present context (Argamon, 2008). In fact, the larger p, the more emphasis is put on entries with large differences, and, as p approaches infinity, only the maximum difference matters, $||x - y||_\infty = \max_{i=1}^{n} |x_i - y_i|$. Rather common choices are $p = 1, 1.4, 2, 4$ (Büttner *et al.*, 2017).

The feature vectors are called vectors since they can be visualized as arrows in the n dimensional space. The δ_p measures above have in common that they focus on the distance of the endpoints of those vectors. A different approach is often used in information retrieval to find documents similar to a query text. With

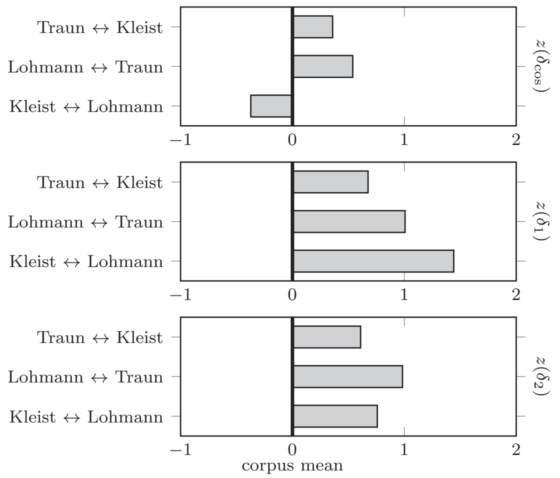


Fig. 8 We compare pairwise distances of the 500 mfw for the three novellas of *Kleist*, *Lohmann*, and *Traun*. To make distances comparable we show the discrepancy to the mean in units of standard deviation (i.e. *z*-scores of distances). For cosine distance, Kleist and Lohmann are closest whereas Lohmann and Traun are far apart. For Burrows's Delta the novellas of Kleist and Lohmann are far apart and Traun and Kleist have the smallest distance. Finally, for quadratic delta, the distance of Lohmann and Traun is larger than the one of Kleist and Lohmann.

$$\mathrm{sim}_{\cos}(x, y) = \frac{1}{||x||_2 \cdot ||y||_2} \cdot \sum_{i=1}^{n} x_i \cdot y_i \ ,$$

we can compute the cosine between the angle of the two vectors and hence, this is called 'cosine similarity'. The reverse, $\delta_{\cos}(x, y) = 1 - \mathrm{sim}_{\cos}(x, y)$, is called 'cosine distance'. It places the focus on the similarly skewed distributions of weighted term frequencies, rather than individual frequencies. If the feature vectors are already normalized, $\delta_{\cos}$ is essentially the same as δ_2 since $||x||_2 = 1 = ||y||_2$ implies $||x - y||_2^2 = 2\delta_{\cos}(x, y)$.[2] The choice of distance measure is not only of quantitative importance, but may lead to qualitatively differnet results (Fig. 8).

Independent of the particular term occurrence-based construction, we can compute the feature vectors $x(D_1), \ldots, x(D_N)$ with documents $D_1, \ldots, D_N$ in the corpus $\mathcal{C}$. With any dissimilarity measure δ defined on them, we obtain a document-dissimilarity matrix $\Delta = (\delta_{k\ell}) \in \mathbb{R}_{\geq 0}^{N \times N}$ with entries $\delta_{k\ell} = \delta(t(D_k), t(D_\ell))$. This matrix summarizes the relationships between the texts in the corpus with respect to the frequencies of words from which they are composed.

3.4 Clustering, scatterplots, and networks

To structure a corpus into groups of texts that are similar within, and dissimilar across groups, a matrix of dissimilarities is constructed as outlined in the previous section and then subjected to a clustering method.

Each clustering method strikes a different balance between the number of groups, group sizes, intra-group similarity, and inter-group dissimilarity. Standard clustering approaches include agglomerative hierarchical clustering, *k*-means, or density-based clustering (DBSCAN). There is an abundance of research on clustering methods for various contexts (Estivill-Castro, 2002; Berkhin, 2006).

Since we do not want the effects of different notions of word occurrence-based dissimilarity measures to be confounded by the selection of a particular clustering method, we refrain from applying one at all. Instead, represent the dissimilarity matrices such that groupings likely to be stable across multiple clustering methods are recognizable.

To represent dissimilarity matrices we choose networks over the more common scatterplots. The reason is explained and illustrated in the following.

Scatterplots are typically obtained from dimensionality reduction methods such as multidimensional scaling (MDS) or a principal component analysis (PCA). They are common for both exploration and presentation of similarity-based clusterings, but pairwise dissimilarities are necessarily distorted when projecting them into only two or three dimensions. Whether this affects the recognizability of clusters depends on the context.

Consider as an example the boxed scatterplots in Fig. 9. All three represent the same δ_2-distance matrix between feature vectors containing (case-sensitive) term frequencies of the 100 most frequent words in the *Novellenschatz* corpus. The PCA scatterplot thus reproduces Fig. 6 from Jannidis (2017) where the group of female authors is found to cluster in the upper left quadrant. This is of interest because PCA maximizes variance one dimension at a time. Female authors are located in a group that is somewhat recognizable in the primary (horizontal) dimension and clearly distinguished along the secondary (vertical) dimension.

The other two scatterplots are obtained from other common dimensionality-reduction methods. In the scatterplot obtained from MDS, axes are not relevant. Positions are determined instead to minimize the percentage error in the representation of dissimilarities by distances in the scatterplot. The MDS scatterplot suggests that female authors are rather peripheral but do not cluster.

PCA and MDS generally provide an overview of similarities and dissimilarities in terms of spatial distance, but may fail to represent clusters well because of overplotting in low-dimensional display space. A more recent non-linear dimensionality-reduction technique, *t*-distributed stochastic neighbor embedding (*t*-SNE, van der Maaten and Hinton 2008), yields a more pronounced clustering overall, and for all but three female authors in particular.

All three methods are designed to minimize misrepresentation of input distances in low-dimensional output space, but their underlying objectives represent different trade-offs between large and small misrepresentations of large and small distances. The focus on nearby nodes in the formulation of *t*-SNE, for instance, is apparent in the result.

Even for apparent groupings, uncertainty remains. Each dimensionality-reduction method represents a unique compromise, and it is difficult to assess whether what appears to be a strong clustering is really a coincidence resulting from that compromise.

3.5 Backbone networks

In an attempt to represent dissimilarities without introducing projection artifacts, we opt for a representation that is more qualitative and retains more degrees of freedom.

A (dis)similarity matrix can be viewed as a complete weighted network, and one way to reduce it to only the strongest similarities is by thresholding. However, a certain level of similarity that may be high for one text can be comparatively low for another, for instance, because the latter is part of a group of mutually similar texts. An alternative, based on relative rather than absolute similarity, are nearest-neighbor networks in which relationships with the most similar other texts are retained.

We here use a restricted variant of Simmelian backbones (Nick *et al.*, 2013), a filtering technique that uses locally adaptive thresholds considering also the vicinity around a pair of texts. It is designed to keep only those pairwise relationships that are relatively strong and sufficiently reinforced by joint similarity to others.

For each novella, the other novellas can be ranked by their similarity to the first, and we consider a fixed number, say ten, of the most similar ones, independent of their absolute similarity score. The networks resulting from this first step are nearest-neighbor networks. They are weighted by similarity rank and also directed, because a novella may be among the most similar of another without the reverse being the case as well.

In a second step, we remove all those relationships where the similarity may be relatively high but not indicative of joint group membership. The relationship of a text with a neighboring text is retained only if it has, among its ten most similar others, at least five of the neighbors of the first text. This results in a tendency to maintain relationships with texts that are not relatively more similar to other groups.

The backbone network in Fig. 9 corroborates the groupings from the PCA and *t*-SNE scatterplots but also indicates that at least one novella with a female author is less strongly linked to the main cluster of female authors. Emmy von Dincklage's *Der Striethast* is located above but close to eight other female-authored novellas in the upper left of the *t*-SNE scatterplot. The backbone network suggests that this is an artifact of the projection because it does not retain any link to that group. Dincklage's novella is a pendant vertex in the top middle of the backbone network and shows strong similarity only to one novella of a male author (Alexander von Sternberg).

The layout of a network is not given but generally determined such that densely connected groups are placed closer together, and unconnected parts farther apart to allow for an interpretation similar to scatterplots. Still, no deeper meaning should be read into the layouts, as the analytic information is in the local structure of links rather than relative positioning of nodes. We use standard network visualization tools and some manual editing to make the structure visible where the layout algorithms do not and to reduce differences in layouts across similar networks.

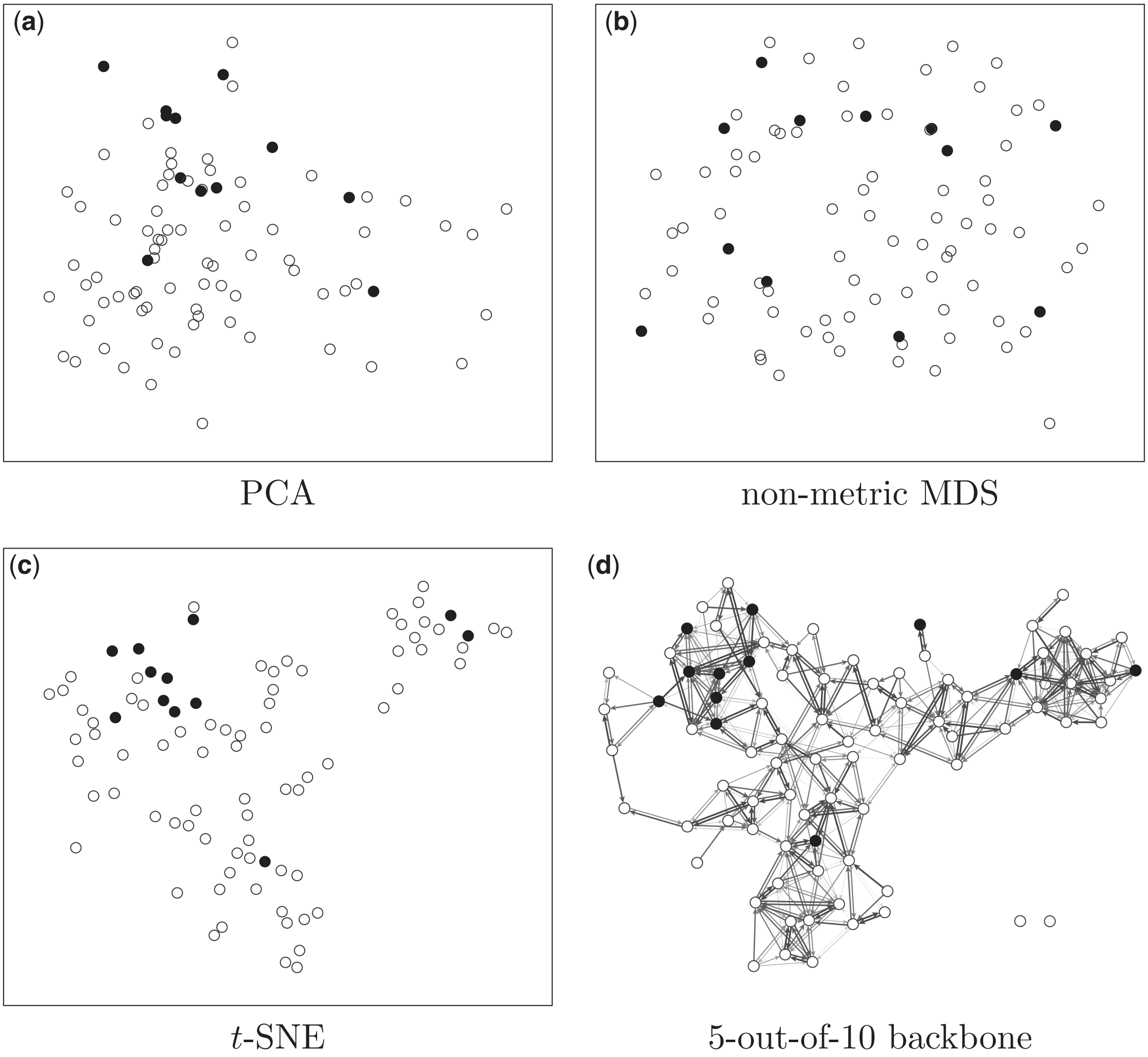


Fig. 9 Three two-dimensional scatterplots and a network backbone representing the (dis)similarities between texts in the *Novellenschatz* based on the 100 most frequent words. Black dots highlight novellas written by female authors. Note that the backbone is a structure with no prescribed geometry; the layout has been determined for clarity but should be considered flexible. (a) PCA; (b) non-metric MDS; (c) *t*-SNE; (d) 5-out-of-10 backbone.

Note that, in principle, we could have used scatterplot positions for layout; and the links would still add information on where the actual similarities are. Despite being close in *t*-SNE coordinates, Dincklage's novella would still be recognized as more dissimilar from the novellas of other female authors because the backbone contains no link between hers and theirs.

4 Demonstrations

A full quantitative analysis of the sensitivity of stylometric similarities with respect to changes in the measures and their parametrization is beyond the scope of this contribution. There are simply too many degrees of freedom, and any distinction is muted or amplified by the specific corpus studied. We instead content ourselves with raising awareness for the non-negligible consequences that choices of features and similarity measures have on corpus analysis.

All examples in this section consist of the eighty-six novellas of the *Novellenschatz* and are based on the frequencies of words that have been converted to lower-caps only. No stemming or stop-word filtering was applied, but we did filter words that appear in only few novellas (20% culling). From the 500 most frequent words thus obtained for the corpus, we generate three feature vectors for each novella. One consists of relative frequencies of words (*tf*-scores) and in the other two they are weighted by their prevalence in the documents (*tf-idf*-scores), and replaced by their normalized deviation from the expectation (*z*-scores). The meta-data used in the analysis is given in Table A1.

To determine the (dis)similarity of novellas, we compare their 500-dimensional feature vectors using absolute differences between their entries (δ_1-distance) as well as Euclidean distance (δ_2-distance) and cosine similarity ($\delta_{\cos}$-distance) in the feature space. This yields nine combinations of feature vectors and dissimilarities.

To understand grouping tendencies that most clustering methods will exhibit, a Simmelian backbone is determined for each of these nine dissimilarity matrices. For each novella, we create a ranking of the ten most similar other novellas. A link is created from one novella to another, if at least five of the ten novellas closest to it are also in the top ten of the other. We also require that the neighbor is among the top ten itself, so that the relation need not be symmetric.

The resulting backbones are shown in Fig. 10. The layouts have been adjusted to ease recognition of similar substructures. While all combinations of feature vectors and distances suggest that there are groups of relatively more similar novellas, substantial differences seem to exist. Without knowing what links them together we have chosen one apparent group in the lower right and highlighted the corresponding novellas in all nine backbones. We will return to this group at the end of the section, but want to discuss first how known groups can be found in some backbones but not others. This serves to demonstrate that, for groups that are not defined by authorship, the choices made during the construction of similarity measures have a strong impact on the potential for uncovering groups using clustering.

In Fig. 11, the subgroup of *Adelsnovellen* (novellas involving nobility) is clearly identifiable, if clustering is based on the presence of associated words of high discriminatory power. With 40% culling, however, the group dissolves (Fig. 12) because the indicative noble titles are no longer part of the feature vector (cf. Fig. 5). Without prior knowledge of the significance of certain words for a subgenre, focusing on more widely appearing words bears the risk of losing the possibility to identify a relatively small group characterized by them.

We can observe very similar consequences with first-person narration. The pronounced cluster on the left of Fig. 13 becomes part of a larger group if we use generic pronouns and lemmatization instead. Note that this may very well be the desired outcome, for instance, if groups of novellas are sought while controlling for the perspective in which they are narrated. Depending on analytic interest, the presence of certain pronouns among the most frequent words may thus be a signal or a confounding factor.

This is further emphasized in Fig. 14, where we determined main characters as those with maximum degree in the co-occurrence network, i.e. characters who are referred to in the same paragraph with the largest number of others. Novellas with a female main character tend to cluster, and these groups are very similar to those obtained when selecting novellas in which `sie` (she) or `die` (female definite article) is the most frequent word.

The examples in Figs 15 and 16 demonstrate that sometimes very few words explain a clustering obtained after a series of steps obfuscating their significance. Cohesive groups in Fig. 15 can be discriminated with a short list of articles, pronouns, and a conjunction, if features are based on term frequencies.

Amplifying the frequency of otherwise unusual words with *tf-idf*-scores, on the other hand, leads to clusters determined by the cast of characters as shown in Fig. 16. The words `Pfarrer`, `Fräulein`, `Bruder`, `Vetter`, and `Graf` are the five nouns with the highest average *tf-idf* -scores across all novellas, and also have large variance.

Of course, the presence or absence of certain groups rarely depends on one single, or even a small class of, features. A cluster that is apparent in one

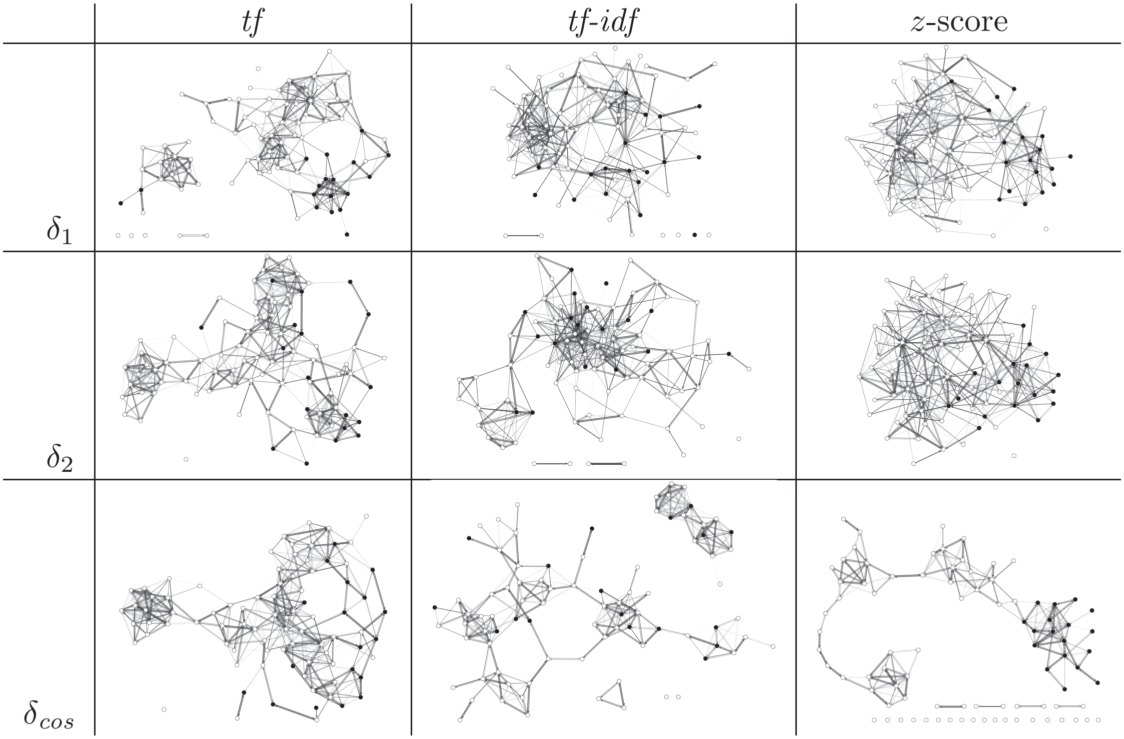


Fig. 10 Backbone networks based on three different distance measures and three different feature vectors from the 500 mfw. An edge points from one novella to another, if five out of the ten novellas closest to the first are also among the ten closest to the second. Edge thickness indicates the rank of the neighbor among the closest novellas. For comparison, we highlighted an apparent group from the lower right (cosine similarity of *z*-scores) in all backbones; this mystery group is discussed at the end of the section.

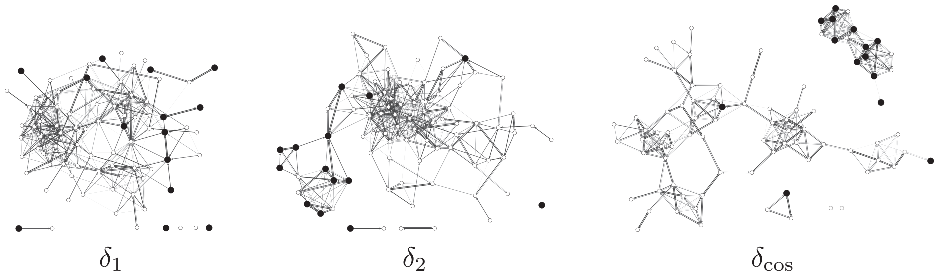


Fig. 11 Backbones based on *tf-idf*-scores where highlighted nodes represent *Adelsnovellen* (novellas involving nobility). They tend to cluster if larger deviations in single entries are emphasized, which δ_1 does not. According to Fig. 15 (middle), the most indicative words for this subgenre are `Graf` (noble title) and `Fräulein` (young lady).

backbone and relatively stable across multiple analyses was already highlighted in Fig. 10. It consists of fourteen novellas written by *Arnim, Halm, Heyden, Gottfried and Johanna Kinkel, Kleist, Kruse, Kugler, Kurz, Lorm, Müller, Raabe, Riehl, Rumohr, Schücking, Sternberg, Waldmüller*, and *Wallner*.

Searching for a reason why these novellas are considered similar across a number of operationalizations, we find that thirteen of the forty verbs most underrepresented in this subgroup[3] are associated with direct speech. Storm (1920) later referred to the novella as 'Schwester des Dramas' (drama's sister),

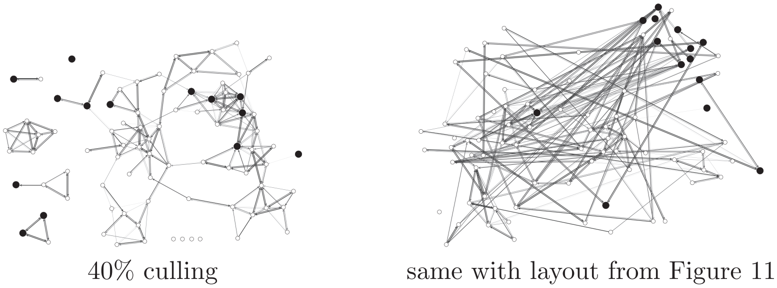


Fig. 12 Backbones based on cosine similarity of *tf-idf*-scores with 40% culling where highlighted nodes represent *Adelsnovellen*. The layout on the right is the same as in Fig. 11 (rightmost) and thus illustrates the large differences between 20% and 40% culling where, e.g. `Graf` is no longer a feature.

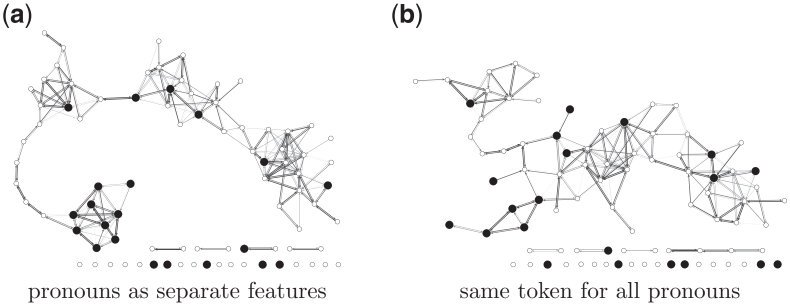


Fig. 13 Backbone networks based on cosine similarity of *z*-scores. One cluster consists entirely of first-person narratives (dark nodes) and is largely due to an overrepresentation of the word `ich` (first person pronoun). Combining pronouns into one generic token has a substantial effect on the cluster structure. (a) Pronouns as separate features and (b) same token for all pronouns.

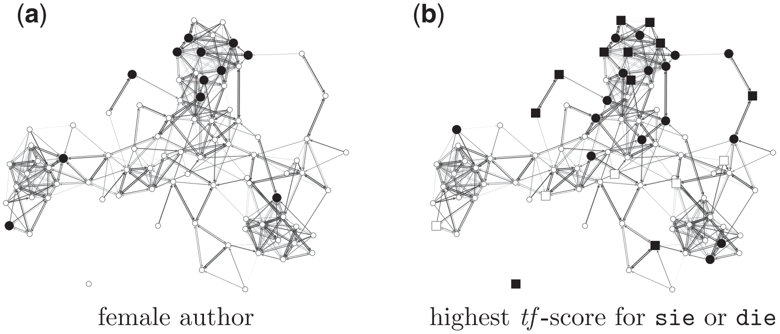


Fig. 14 Backbone network based on δ_2 of *tf*-scores. Female authors tend to cluster (dark nodes on the left) as do novellas (dark nodes on the right) in which the most frequent word is either `sie` (she) or `die` (female definite article). These are not the same as the novellas with a female main character (squares). (a) Female author and (b) highest *tf*-score for `sie` or `die`.

which prompts us to expect a prevalence of direct speech. A straightforward test for the presence of direct speech is challenging because of the variety of markers used; instead of by quotation marks, direct speech is often indicated by starting a new line, hyphens, or no syntactical element at all.

A second factor that appears to contribute to the discrimination of this group is the more frequent use

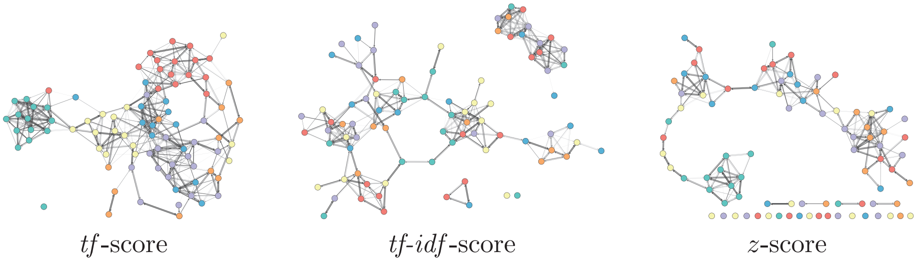


Fig. 15 Backbones for cosine similarity of the 500 mfw. Each novella is colored according to the word with highest *tf*-score among `er` 🔵, `der` 🟣, `sie` 🔴, `die` 🟠, `ich` 🟢, and `und` 🟡. For the *tf*-features, but not the others, most of the clustering is already explained by this one word. A color version of this figure appears in the online version of this article.

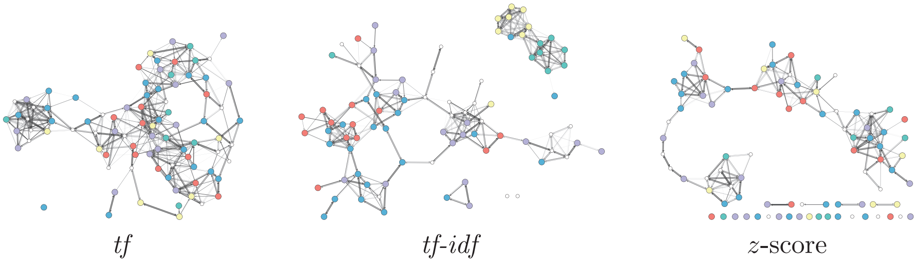


Fig. 16 Backbones for cosine similarity with novellas colored according to the word with highest *tf-idf*-score among `Pfarrer` 🔴, `Fräulein` 🟡, `Bruder` 🔵, `Vetter` 🟣, and `Graf` 🟢. The other novellas contain neither of these words. For the *tf-idf*-features, but not the others, most of the clustering is already explained by this one word. A color version of this figure appears in the online version of this article.

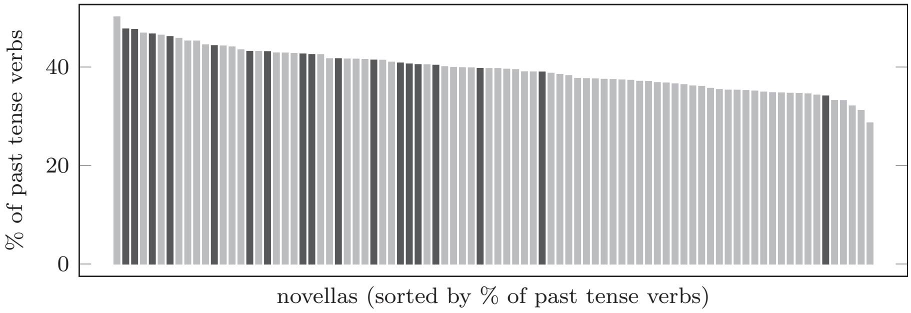


Fig. 17 With the exception of the novella by *Heyden*, those in the mystery group from Fig. 10 (marked black) have a comparatively high use of past tense verbs.

of past tense.[4] As shown in Fig. 17, all but two are above the median.

5 Methodological Consequences

We have broken down the process of grouping texts by stylometric features, and more specifically the similarity of word use. Using the corpus of *Deutscher Novellenschatz* as a case study, we have demonstrated that the choices made in the grouping process have substantial impact on the groups found, and that groups are sometimes determined by seemingly trivial factors which may or may not be desirable for the research question at hand.

As a consequence, discussions of the suitability of similarity measures need to take into account the context in which the measures are applied, how the data are prepared, and which kind of signal is discriminative. Author signals, genre characteristics, narrative perspective, plot elements, and many other aspects confound the definition of stylometric similarity. The task of amplifying desired aspects and controlling for others is thus both a theoretical and an empirical one, and it should not be taken lightly.

If a resulting grouping matches the expectation, but the employed stylometric similarity does not operationalize the hypothesized mechanism, this still does not confirm a grouping hypothesis. Confidence in the meaningfulness of a clustering is gained only through justified decisions, not from decisiveness of results or their robustness with respect to different parameters.

For a given corpus and an idea about distinctive qualities of the texts in it, the task is to exploit the individual steps in the clustering process to let it be governed by these qualities. If individual authors can be distinguished by their use of function words, but male–female differences are suspected to manifest themselves in the use of certain adjectives, different stylometric similarity measure should be used for these tasks. We have pointed out a number of options for adaptation, from tokenization to baseline distributions, but the list of course goes on.

Our discussion was focused on the construction of similarities, and did not include the influence of specific clustering methods. There is a qualitative difference between these two aspects, and there is no point in accurate clustering if the similarity measure is inappropriate. This motivated the use of backbone networks because they point to relatively cohesive groups any clustering method will tend to preserve.

Through the presentation of several examples on a medium-size literary corpus, the *Deutscher Novellenschatz*, we attempted to make the practical relevance of these general methodical considerations more tangible. While plagiarism and authorship attribution may be major use cases for stylometric similarity analyses, these examples show that there is great potential for the identification of other clusters of texts, not related to authorship. Each kind of clustering will require its own combination of methods, because what works for authorship attribution may not apply to gender differentiation, and what works for genre classification may not apply to narration styles. Rather than contributing a solution, it seems, we are thus creating more problems.

Funding

This work was supported by Volkswagen Foundation via Project 'Reading at Scale' (Az. 91 920) within the funding program 'Mixing Methods. An Agenda for the Humanities in the Digital Age'.

Appendix

Table A1. Used metadata in the analysis of the novellas

Author	Title	First-person narrative	Gender	*Adelsnovelle*
Alexis, W.	Herr von Sacken		m	x
Andolt, E.	Eine Nacht	x	m	x
Arnim, A. v.	Der tolle Invalide auf dem Fort Ratonneau		m	
Auerbach, B.	Die Geschichte des Diethelm von Buchenberg		m	
Berthold, F.	Irrwisch-Fritze		f	
Brentano, C.	Geschichte vom braven Kasperl und dem schoenen Annerl	x	m	x
Chamisso, A. v.	Peter Schlemihl's wundersame Geschichte	x	m	
Dincklage, E. v.	Der Striethast		f	
Droste-Hüllshof, A. v.	Die Judenbuche		f	
Eichendorff, J. v.	Die Glücksritter		m	x

(*Continued*)

Table A1 (continued)

Author	Title	First-person narrative	Gender	*Adelsnovelle*
Ense, K. A. V. v.	Reiz und Liebe	x	m	
Fräulein v. Wolf	Gemüth und Selbstsucht		f	
Frey, J.	Das erfüllte Versprechen		m	x
Gall, L. v.	Eine fromme Lüge		f	x
Gerstäcker, F.	Germelshausen		m	
Glümer, C. v.	Reich zu reich und arm zu arm		f	
Goethe, J. W. v.	Die neue Melusine	x	m	
Goldammer, L.	Eine Hochzeitsnacht		m	
Goldammer, L.	Auf Wiedersehen!		m	
Gotthelf, J.	Kurt von Koppigen		m	
Gotthelf, J.	Der Notar in der Falle		m	
Grillparzer, F.	Der arme Spielmann	x	m	
Grimm, H.	Das Kind		m	
Gross, J.	Vetter Isidor		m	x
Hackländer, F. W.	Zwei Nächte		m	x
Halm, F.	Die Marzipan-Lise		m	
Hartmann, M.	Das Schloß im Gebirge	x	m	
Hauff, W.	Phantasien im Bremer Ratskeller	x	m	
Heyden, F. v.	Der graue John		m	
Heyse, P.	Der Weinhüter von Meran		m	
Höfer, E.	Rolof, der Rekrut		m	
Hoffmann, E. T. A.	Das Fräulein von Scuderi		m	
Holtei, K. v.	s Muhme-Leutnant-Saloppel		m	
Horner, H.	Der Säugling		m	
Immermann, K.	Der Carneval und die Somnambule	x	m	x
Kähler, L. A.	Die drei Schwestern	x	m	
Keller, G.	Romeo und Julia auf dem Dorfe		m	
Kinkel, G.	Margret		m	
Kinkel, J.	Musikalische Orthodoxie		f	x
Kleist, H. v.	Die Verlobung von St. Domingo		m	
Kompert, L.	Eine Verlorene		m	
Kopisch, A.	Ein Karnevalsfest auf Ischia		m	
Kopisch, A.	Der Träumer		m	
Kruse, L.	Nordische Freundschaft		m	
Kürnberger, F.	Der Drache		m	
Kugler, F.	Die Incantada		m	
Kurz, H.	Die beiden Tubus		m	
Lewald, F.	Die Tante	x	f	
Lohmann, F.	Die Entscheidung bei Hochkirch		f	
Lorm, H.	Ein adeliges Fräulein	x	m	x
Ludwig, J.	Das Gericht im Walde		f	
Meißner, A.	Der Müller vom Höft		m	
Meyr, M.	Der Sieg des Schwachen		m	
Mörike, E.	Mozart auf der Reise nach Prag		m	x
Mügge, T.	Am Malanger Fjord		m	
Müller, W.	Debora		m	
Pichler, A.	Der Flüchtling	x	m	
Raabe, W.	Das letzte Recht		m	
Reich, M.	Mammon im Gebirge		m	
Riehl, W. H.	Jörg Muckenhuber		m	
Roquette, O.	Die Schlangenkönigin	x	m	
Rumohr, K. F.	Der letzte Savello		m	
Sacher-Masoch, L.	Don Juan von Kolomea		m	

(*Continued*)

Table A1 (continued)

Author	Title	First-person narrative	Gender	*Adelsnovelle*
Schefer, L.	Die Düvecke, oder die Leiden einer Königin		m	
Scheffel, J. V. v.	Hugideo		m	
Schmid, H.	Mohrenfranzel	x	m	
Schreyvogel, J.	Samuel Brinks letzte Liebesgeschichte	x	m	
Schücking, L.	Die Schwester		m	x
Spindler, K.	Die Engel-Ehe		m	
Sternberg, A. v.	Scholastika		m	
Stifter, A.	Brigitta	x	m	
Storm, T.	Eine Malerarbeit	x	m	
Tesche, W.	Der Enten-Piet		m	
Tieck, L.	Die Gemälde		m	
Tieck, L.	Des Lebens Überfluss		m	
Traun, J. v. d.	Der Gebirgspfarrer		m	
Waldmüller, R.	Es ist nicht gut, daß der Mensch allein sei		m	
Wallner, F.	Der arme Josy	x	m	
Wichert, E.	Ansas und Grita	x	m	
Widmann, A.	Die katholische Mühle	x	m	
Wilbrandt, A.	Johann Ohlerich		m	
Wild, H.	Eure Wege sind nicht meine Wege		f	x
Wildermuth, O.	Streit in der Liebe und Liebe im Streit		f	
Wolf, A.	Der Stern der Schönheit		m	
Ziegler, F. W.	Saat und Ernte		m	
Zschokke, H.	Der tote Gast		m	

References

Argamon, S. (2008). Interpreting Burrows's delta: geometric and probabilistic foundations. *Literary and Linguistic Computing*, **23**(2): 131–47.

Baeza-Yates, R. and Ribeiro-Neto, B. A. (2011). *Modern Information Retrieval - The Concepts and Technology Behind Search*, 2nd edn. Harlow: Pearson Education Ltd.

Berkhin, P. (2006). *A Survey of Clustering Data Mining Techniques*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 25–71.

Büttner, A., Dimpel, F. M., Evert, S., Jannidis, F., Pielström, S., Proisl, T., Reger, I., Schöch, C., and Vitt, T. (2017). "Delta" in der stilometrischen Autorschaftsattribution. *Zeitschrift für digitale Geisteswissenschaften*, **Heft 2**: Artikel 6.

Burrows, J. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**(3): 267–87.

Burrows, J. (2007). All the way through: testing for authorship in different frequency strata. *Literary and Linguistic Computing*, **22**(1): 27–47.

Estivill-Castro, V. (2002). Why so many clustering algorithms: a position paper. *SIGKDD Explorations Newsletter*, **4**(1): 65–75.

Jannidis, F. (2017). Perspektiven quantitativer Untersuchungen des Novellenschatzes. *Zeitschrift für Literaturwissenschaft und Linguistik*, **47**(1): 7–27.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Scoring, Term Weighting, and the Vector Space Model*. Cambridge: Cambridge University Press, pp. 100–23.

Nick, B., Lee, C., Cunningham, P., and Brandes, U. (2013). Simmelian backbones: amplifying hidden homophily in Facebook networks. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, Niagara, Ontario, Canada, August 25–28, 2013, pp. 525–32.

Schöch, C., Schlör, D., Zehe, A., Gebhard, H., Becker, M., and Hotho, A. (2018). Burrows' zeta: exploring and evaluating variants and parameters. In Palau, J. G. and Russell, I. G. (eds), *13th Annual International Conference of the Alliance of Digital Humanities Organizations (DH 2018)*, Mexico City, Mexico, June 26–29, 2018, pp. 274–77.

Schöch, C. (2014). Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik. In Schneide, C. S. L. (ed.), *Literaturwissenschaft im digitalen Medienwandel*, Beihefte zu Philologie im Netz. PhiN. Licence Creative Commons Attribution 4.0 (CC-BY), pp. 130–57.

Sidorov, G., Gelbukh, A., Gómez-Adorno, H., and Pinto, D. (2014). Soft similarity and soft cosine measure: similarity of features in vector space model. *Computación y Sistemas*, **18**(3): 491–504.

Smith, P. W. and Aldridge, W. (2011). Improving authorship attribution: optimizing Burrows' Delta method. *Journal of Quantitative Linguistics*, **18**(1): 63–88.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, **60**(3): 538–56.

Storm, T. (1920). Eine zurückgezogene Vorrede aus dem Jahre 1881. In Köstner, A. (ed.), *Theodor Storm, Sämtliche Werke*, vol. **8**. Leipzig: Insel-Verlag, pp. 122–23.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, **9**: 2579–2605.

Weitin, T. (2016). Selektion und Distinktion. Paul Heyses und Hermann Kurz' Deutscher Novellenschatz als Archiv, Literaturgeschichte und Korpus. In *Archiv/Fiktionen. Verfahren des Archivierens in Literatur und Kultur des langen 19. Jahrhunderts*, pp. 385–408.

Weitin, T. (in press). Burrows's Delta und Z-Score-Differenz im Netzwerkvergleich. Analysen zum Deutschen Novellenschatz von Paul Heyse und Herrmann Kurz (1871–1876). In Jannidis, F. (ed.), *Digitale Literaturwissenschaft: DFG-Symposion 2017*. Stuttgart: Metzler.

Weitin, T. and Herget, K. (2017). Falkentopics: Über einige Probleme beim Topic Modeling literarischer Texte. *Zeitschrift für Literaturwissenschaft und Linguistik*, **47**(1): 29–48.

Zipf, G. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison–Wesley.

Zipf, G. K. (1935). *The Psychobiology of Language*. New York: Houghton-Mifflin.

Notes

[1] If we just consider terms i that appear somewhere in the corpus, we can assure that $N(i) > 0$.

[2] A generalization of cosine similarity has been proposed in Sidorov *et al.* (2014). They suggest to use an additional term-similarity matrix $S = (s_{ij}) \in \mathbb{R}_{\geq 0}^{n \times n}$ to weight the contributions of all pairs of entries x_i, y_j. This *soft cosine* similarity measure

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}x_i y_j}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}x_i x_j}\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}y_i y_j}}$$

reduces to cosine similarity for $S = I_n$, i.e.,

$$s_{ij} = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j, \end{cases}$$

where terms are similar only to themselves. Similarities can be motivated syntactically or semantically, and thus allow for a more nuanced aggregation than the combination of multiple words into a single term feature.

[3] The verbs with the least average z-score in this group are: sind, ist, **sagte**, bin, ging, thun, muß, bist, **gesagt**, fuhr, kannst, kommt, **glaube**, will, **rief**, geht, **denken**, hast, kommen, **hörte**, sehe, thut, **gehört**, **sagen**, kann, **weißt**, **dachte**, **weiß**, **wissen**, lachte, willst, sah, gehen, hat, helfen, gethan, **reden**, essen, saßen, wird.

[4] Classification of tense is notoriously difficult for German. We used tags from the German-language model of spacy (https://spacy.io/models/de) and decided for past tense as follows:

```
Input word.tag and word.text from spacy and tense
if word.text ∈ {war, warst, waren, wart} then
    tense ← past
else if word.tag ∈ {VAFIN, VMFIN, VVFIN} then
    if last two letters = te or = ten then
        tense ← past
    else
        tense ← present
    end if
else if word.tag ∈ {VAPP, VMPP, VVPP} then
    tense ← past
else if word.tag ∈ {VAIMP, VAINF, VMINF, VVIMP, VVINF, VVINF, VVIZU} then
    tense ← present
end if
```