Reading at Scale

A Digital Analysis of German Novellas from the 19th Century

Book Chapter

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

Publication date:
2023-10

Permanent link:
https://doi.org/10.3929/ethz-b-000635998

Rights / license:
Creative Commons Attribution 4.0 International

Originally published in:
Digital Humanities Research 7, https://doi.org/10.1515/9783839469132-008

Reading at Scale. A Digital Analysis of German Novellas from the 19th Century (Reading at Scale)

Thomas Weitin, Simon Päpcke, Katharina Herget, Anastasia Glawion, Ulrik Brandes

> “das Beste, was in dieser Gattung geleistet ist [...] zu sammeln und in übersichtlicher Folge herauszugeben, bedarf wohl kaum der Rechtfertigung“.[1]

Abstract *The Deutscher Novellenschatz (published in 24 volumes 1871–1876) is a collection of 86 German-language novellas edited by Paul Heyse and Hermann Kurz. It is an example of a medium-sized corpus amenable, in principle, to both, scholarly reading and automated analysis. Our point of departure was the conviction that research questions at intermediate granularity would require the combination of hermeneutic and statistical methods to achieve an appropriate level of abstraction while maintaining a sufficient amount of context. By exposing the sensitivity of text similarity measures to choices in the preparation and evaluation of bag-of-word representations, we highlight the need for consideration of contextual information even in the most distant reading approaches. Literary theory suggested more coarse-grained hypotheses that we tested both empirically in a group of non-expert readers and computationally using similarity of character constellations, respectively. Based on correspondences of the editors and comparison with other corpora, the Novellenschatz was further situated in a historiographic context.*

Introduction

An important yet rarely studied issue in Digital Philology as a subdomain of the Digital Humanities is the composition of text corpora. Despite the availability of vari-

[1] “To collect the best that has been achieved in this genre [...] and to publish it in a well arranged order hardly needs justification”, translated by the authors. Paul Heyse and Hermann Kurz, “Einleitung”, in *Deutscher Novellenschatz*, vol. 1 (München: R. Oldenbourg, 1871).

ous repositories,[2] there is no gold standard for corpus composition in digital literary studies. Such a standard is perhaps unattainable, because literary text corpora must be prepared individually and purposefully according to the specific research question of the project at hand. Even more so, the constitution of digital literary corpora has to be critically reflected upon.

Within the 'Reading at Scale' project, we focused on the novella collection *Deutscher Novellenschatz*. One of the distinctive features of this research object is that it is both a corpus and an artefact at the same time: With its overall 213 novellas, the *Novellenschatz*-series—including the *Deutscher Novellenschatz* (DNS), the *Neuer Deutscher Novellenschatz* (NDNS) and the *Novellenschatz des Auslandes* (NSdA)[3]—comprises a number of popular German-language novellas (and in the case of the *NSdA* foreign-language novellas translated into German) suitable for statistical analysis.[4] As an artefact, the *Novellenschatz*-collection is an object of historical interest, specifically with regard to its normative claims showcased in the quotation above. The novella collection is complemented by letters that the main editor Paul Heyse exchanged with the others: Hermann Kurz until his passing in 1873 and Ludwig Laistner thereafter. In the letters, the editors thoroughly discuss some of the choices they made for the collection. This correspondence, which is unfortunately only published in excerpts,[5] highlights that the notion of relationality was of particular relevance for the selection process.

The so-called long 19th century (1789–1914) is well known for its rapidly growing mass market for literature, including magazines and newspapers.[6] In addition, the novella—especially the Realistic novella—became the epitome of this mass production in the German-speaking literary world. This progressive development cul-

[2] For example: "Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. Herausgegeben von der Berlin-Brandenburgischen Akademie der Wissenschaften", http://www.deutschestextarchiv.de/, or *Textgrid*, TextGrid Consortium. 2006–2014. TextGrid: A Virtual Research Environment for the Humanities. Göttingen: TextGrid Consortium. textgrid.de.

[3] Paul Heyse and Hermann Kurz (ed.), *Deutscher Novellenschatz* (1871–1876, Bd. 1–24, 86 novellas) München: R. Oldenbourg. Paul Heyse and Ludwig Laistner (ed.), *Neuer deutscher Novellenschatz* (1884–1887, Bd. 1–24, 70 novellas) München: R. Oldenbourg. Paul Heyse and Hermann Kurz (ed.), *Novellenschatz des Auslandes* (1877–1884, Bd. 1–14, 57 novellas) München: R. Oldenbourg.

[4] The collection is completely digitized and is already being published (Weitin 2016; 2018).

[5] Monika Walkhoff, *Der Briefwechsel Zwischen Paul Heyse Und Hermann Kurz in Den Jahren 1869–1873 Aus Anlass Der Herausgabe Des Deutschen Novellenschatzes [Mit Faks.]* (München: Foto-Druck Frank, 1967).

[6] Matt Erlin and Lynne Tatlock, "Introduction: 'Distant Reading' and the Historiography of Nineteenth Century German Literature", in *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, ed. Matt Erlin and Lynne Tatlock, Studies in German Literature, Linguistics, and Culture (Rochester, New York: Camden House, 2014), 1–29.

minates in the *Novellenschatz*-approach "to collect the best that is produced in this genre […] and to publish it in a clear sequence".[7] By setting this requirement, the collection not only reflects the contemporary perception of a literary deluge ('Literaturschwemme') but also responds to it with the programmatic preface that includes the falcon theory ('Falkentheorie'), an instrument to assess the quality of Realistic novellas[8] which led to high esteem for Heyse as a theorist of the novella genre accorded by literary historian Oskar Walzel.[9]

The requirement proposed by the editors mirrors the standards of literary quality established by 19th-century literary historiography, whereby the quality of the individual text is defined in relation to the epoch, its ideals and predecessors.[10] Heyse and Kurz demonstrate it by opening the *Novellenschatz*-series with Goethe's fairytale *Die Neue Melusine* to promote their underlying Realistic agenda in contrast to the classic example of Romanticism.[11] Heyse contextualizes this choice as follows:

> "In these first volumes we must carefully avoid offending the big bunch and save our caviar for the middle of the table, where they have already learned to swallow all kinds of things."[12]

The quote demonstrates that the selection of novellas, their order and position within the collection are strongly affected by their relation to other texts. This is not a notion that is distinctive for the *Novellenschatz*, but rather a constant underlying feature of the literary market that writers of the 'Literaturschwemme' had to take into account, as they were writing with the awareness of broad similarity.[13] Consid-

[7] Paul Heyse and Hermann Kurz, "Einleitung", in *Deutscher Novellenschatz*, vol. 1 (München: R. Oldenbourg, 1871), https://www.deutschestextarchiv.de/heysekurz_einleitung_1871, translated by the authors.

[8] See Thomas Weitin and Katharina Herget, "Falkentopics: Über Einige Probleme Beim Topic Modeling Literarischer Texte", *Zeitschrift Für Literaturwissenschaft Und Linguistik* 47, no. 1 (2017): 29–48, https://doi.org/10.1007/s41244-017-0049-3.

[9] Wilhelm Scherer and Oskar Walzel, *Geschichte Der Deutschen Literatur* (Berlin: Askanischer Verlag, 1921).

[10] E.g. Oskar Walzel, *Die Deutsche Dichtung Seit Goethes Tod* (Berlin: Askanischer Verlag, 1919); Wilhelm Scherer, *Geschichte Der Deutschen Litteratur*, 3. Auflage (Berlin: Weidmannsche Buchhandlung, 1885); Georg Gottfried Gervinus, *Handbuch Der Geschichte Der Poetischen National-Literatur Der Deutschen*, 3. Aufl. (Leipzig: Engelmann, 1844).

[11] See Thomas Weitin, *Digitale Literaturgeschichte. Eine Versuchsreihe in 7 Experimenten* (Berlin: Metzler/Springer Nature, 2021).

[12] "Wir müssen gerade in diesen ersten Bänden sorgfältig vermeiden dem großen Haufen vor den Kopf zu stoßen und unsern Caviar lieber für die Mitte der Tafel sparen, wo sie schon allerlei schlucken gelernt haben". Unpublished correspondence letter, P. Heyse to H. Kurz, 14.10.1870, translated by the authors.

[13] Thomas Weitin, "Average and Distinction. The Deutsche Novellenschatz Between Literary History and Corpus Analysis", LitLab Pamphlet, 6 (2018): 1–23.

ering the influence of text relationality on novella production and on the selection processes for the collections in question, we were particularly interested in methods that reflected these connections. Thus, this publication first presents our thoughts on issues of comparability within a literary corpus in general, and in the second part delves into different preprocessing techniques that affect groupings within the first *Novellenschatz*-collection.

Between contextualization and commensuration

When Franco Moretti coined the term 'distant reading' 20 years ago, he certainly did not foresee its subsequent evolution. Originally, Moretti used it to promote his idea to discover the great unread of world literature: in opposition to close reading, distant reading should transgress the limits of the Western canon by using synthesis/meta-analysis "without a single direct textual reading".[14] Regardless of the original meaning, the term was quickly adopted by digital literary scholars. Although Moretti later admitted that the term was originally meant as a joke,[15] its reception certainly influenced the foundation of the Stanford Literary Lab in 2010. Building upon this development, Martin Mueller introduced the term 'scalable reading' as a "happy synthesis of 'close' and 'distant' reading",[16]which allows one to "zoom" between single text and corpus level through digital surrogates.[17]

As a result of our own corpus analyses, we find the scaling metaphor to be misleading. While we do share Mueller's assertion that the "typical encounter with a text is through a surrogate",[18] no matter whether this means reading a single book edition or analysing the data of an entire corpus, we feel that it is limiting to assume the existence of dimensions along which surrogates can be transformed into one another via scaling. From our interdisciplinary experience as professional readers of literature and analysts of data, it is neither enough to state that "the two scales of analysis [...] need to coexist",[19] nor is it helpful or even necessary to expect the different surrogates to correspond to the levels of detail in the same representation.

[14] Franco Moretti, "Conjectures on World Literature", *New Left Review* 1 (2000): 57.

[15] Franco Moretti, "Conjectures on World Literature", in *Distant Reading* (London; New York: Verso, 2013), 43–62.

[16] Martin Mueller, *Scalable Reading*, 2019, https://scalablereading.northwestern.edu/.

[17] Thomas Weitin, "Thinking Slowly. Reading Literature in the Aftermath of Big Data", *LitLab Pamphlet* #1 (2015): 1–19.

[18] Martin Mueller, "Morgensterns Spectacles or the Importance of Not-Reading", Northwestern University Digital Humanities Laboratory (blog), 2013, https://sites.northwestern.edu/nudhl/?p=433.

[19] Matthew L. Jockers, *Macroanalysis: Digital Methods and Literary History*, Topics in the Digital Humanities (Urbana, IL: University of Illinois Press, 2013), 9.

Instead, we posit that appropriate scales and representations are determined by research question, research design, and project resources. In addition to reading from up close or at a distance, it may be useful to have a variety of angles and lenses at one's disposal. In other words, different representations, or surrogates, may be required even at the same level of abstraction, especially because adjacent levels are generally related in ways that are more complex than simply a change of resolution.

With reading at scale, the methodological problems at stake are the opposing actions of contextualization and commensuration. To understand the subtleties of a case, context helps to distinguish it from others. Some aspects of literature, however, become comparable only if specific features are selected in which they can be treated as being of the same kind, although maybe in different ways.[20] In annotation projects and workflows such as CATMA, this logic is applied to the level of text passages within a singular text.[21] The balanced consideration of discriminating peculiarities and shared features governs what makes an adequate representation and ultimately determines the range of methods available. The specific interpretation of the term contextualization that is relevant here will become more apparent after we relate commensurability to statistical analysis.

Any form of statistical data analysis relies on variables or the assignment of values in a specified range to each entity from a defined domain in order to express levels or qualities of a property they all have in common. Variables are productive only if there is at least the potential for the entities to share the property present in them. A categorical variable encoding the author of a text or a numerical one encoding the number of occurrences of the word 'time' is potentially informative, but a binary variable encoding whether a text starts with a specific phrase is generally not unless the phrase is 'Once upon a time' and the corpus contains fairy tales. For any given text, references to biographic information or past works of an author are highly contextual and rarely suited for variable-based comparison across a corpus.

An example of a variable commonly defined for a literary corpus such as ours is the original publication date of each text. Innocent as this may seem, using such a variable, for example to report the distribution of texts over time, begets a number of assumptions. By referring to the elements of a corpus indiscriminately as texts, we are already suggesting that they are entities of the same kind, or are commensurate. In some sense they are so, because all of them similarly occupy a number of pages in

[20] Bettina Heintz, "Numerische Differenz. Überlegungen zu einer Soziologie des (quantitativen) Vergleichs / Numerical Difference. Toward a Sociology of (Quantitative) Comparisons", *Zeitschrift für Soziologie* 39, no. 3 (1 January 2010), https://doi.org/10.1515/zfsoz-2010-0301.

[21] Evelyn Gius et al., "CATMA 6", 6 April 2022, https://doi.org/10.5281/ZENODO.1470118; Jan Horstmann, "Undogmatic Literary Annotation with CATMA. Functions, Differentiation, Systematization", in *Annotations in Scholarly Editions and Research*, ed. Julia Nantke and Frederik Schlupkothen (De Gruyter, 2020), 157–76, https://doi.org/10.1515/9783110689112-008.

one of the volumes of the *Novellenschatz*. For a different purpose, considering titles with different text lengths, such as Wolf's *Stern der Schönheit* (around 20,000 characters) and Auerbach's *Diethelm von Buchenberg* (400,000 characters), as similar entities may be questionable. And even when it appears appropriate to treat them as being of comparable type, it is a separate, substantive question whether their publication dates constitute a feature by which they can be compared. The differences between periods of writing, contemporaneous developments, versions of a text, or forms of publication may be too stark to allow a meaningful ordering of texts by publication date, let alone an interpretation of the length of time intervals in-between.

Variables such as publication date, genre, or gender of the author are extrinsic to the text and often referred to as metadata.[22] Technically, metadata are data about data, which implies that texts are considered data themselves. Intrinsic, or text-immanent, variables may associate each text in a corpus with, for instance, a word-frequency vector, linguistic indices, or plot complexity. They are generally based on an intermediate representation that is defined for each text individually, and then summarized into corpus-level variables. A prominent example are character constellations, which represent relationships such as co-occurrences in a scene between characters that are specific to a text.[23] The network variables denoting the co-occurrence of a pair of characters in a scene are different for each text. Since their dramatis personae are different, they do not assign values to the same pairs of characters, but represent each text in its own specific way. Characteristics such as an index of centralization for co-occurrence networks, however, are comparable across texts and therefore similar to other text descriptors.

Intermediate representations such as character constellations focus on particular aspects of a text and, therefore, filter out details and abstract more general features from the specificities of a text, but they do so for each text individually. Comparability arises from the shared structure of these text-specific representations. The crucial decision thus lies in the level of detail that, on the one hand, needs to be preserved to sustain important distinctions between texts and that, on the other, needs to be abstracted from to allow comparisons across a corpus. There is, however, no inherent relation between intermediate representations, or partial abstractions, of texts that would correspond to a notion of scaling. Bag-of-words representations, character networks, and event sequences surely represent a text on different levels

[22] Matteo Lorenzini, Marco Rospocher, and Sara Tonelli, "On Assessing Metadata Completeness in Digital Cultural Heritage Repositories", *Digital Scholarship in the Humanities* 36, no. Supplement 2 (5 November 2021): 182–88, https://doi.org/10.1093/llc/fqab036; Christof Schöch, "Big? Smart? Clean? Messy? Data in the Humanities", *Journal of Digital Humanities* 2 (2013): 1–12.

[23] See Benjamin Krautter et al., "Eponymous Heroes and Protagonists – Character Classification in German-Language Dramas", *LitLab Pamphlet* #7 (2018): 58, https://www.digitalhumanitiescooperation.de/en/pamphlete/pamphlet-7-interpretierbare-figurenklassifikation/.

of abstraction, but neither of these levels refines or coarsens another. The appropriate choice of scale is determined substantively by the research question, but also pragmatically by the possibility for objective and manageable realization.

Reading at scale is thus fundamentally about identifying a suitable combination of qualitative and quantitative aspects to determine representations, or mixing methods. In data analysis, the qualitative and quantitative notions often refer to the level of measurement of a variable. Nominal and ordinal variables have values representing (un)ordered categories. They do not admit arithmetical operations and are therefore considered qualitative data, whereas values on a ratio-scale express multiples of a unit and are therefore quite literally quantitative. Consequently, there can be qualitative and quantitative variables, and the distinction is made based on properties of the range of values. Since literary epochs generally do not progress linearly, one might argue that publication dates are on an ordinal level of measurement at most.

In hermeneutics, the discomfort with quantitative analyses appears to occur even earlier: often, the very assumption that information can be represented meaningfully in variables is resented. In other words, the commensurability of entities is questioned because every such attempt would rid the subject of crucial circumstantial evidence.

Every mixed-method approach faces this kind of trade-off, and we can work from both ends to arrive at a scale.[24] If we decide to compare the single texts of our corpus only through vectors of word frequencies, the bag-of-words model can be enriched with metadata so that the analysis of historical context becomes a question of subsetting. In Natural Language Processing, word embeddings can be used to foster context sensitivity with respect to both semantics and syntax.[25] And even the results of stochastic semantics as in topic models can be recontextualized through concordance analyses.[26] On the other side, the century-long history of modern philology provides us with resilient methods to avoid drowning in the contextual associations

[24] Rabea Kleymann, "Datendiffraktion: Von Mixed zu Entangled Methods in den Digital Humanities", HTML,XML,PDF, ed. Manuel Burghardt et al., *Fabrikation von Erkenntnis – Experimente in den Digital Humanities (Zeitschrift für digitale Geisteswissenschaften / Sonderband)*, 2022, https://doi.org/10.17175/SB005_008; Andrew Piper, *Enumerations: Data and Literary Study* (Chicago, IL; London: The University of Chicago Press, 2018).

[25] See Simon Hengchen et al., "A Data-Driven Approach to Studying Changing Vocabularies in Historical Newspaper Collections", *Digital Scholarship in the Humanities* 36, no. Supplement 2 (5 November 2021): 109–26, https://doi.org/10.1093/llc/fqab032; Sahar Ghannay et al., "Word Embedding Evaluation and Combination", in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (LREC 2016, Portorož, Slovenia: European Language Resources Association (ELRA), 2016), 300–305, https://aclanthology.org/L16-1046.

[26] See VinhTuan Thai and Siegfried Handschuh, "Context Stamp: A Topic-Based Content Abstraction for Visual Concordance Analysis", in *Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems – CHI EA '11* (the 2011 annual confer-

of close reading. For every author, research proves some key passages to be more relevant than others. Depending on new developments, habits, and theoretical umbrella terms in the humanities, neglected parts of literary works become the focus of new readings and interpretations. Within that framework of innovation, a close reading as the dissociative reading of excerpts ('Stellenlektüre') works only as much as a reduction of possible contexts as the apparatus of secondary literature does, to which scholars turn first when they start to explore new research areas.

Stylometric text similarities revisited

As a concrete example on the highest level of context reductions, we next discuss stylometric similarity in the context of corpus analysis. This section draws heavily on our recent study within the scope of the *Novellenschatz,*[27] where many more aspects and detailed examples can be found.

It is notable that even at this almost extreme end of distant reading, many opportunities exist for contextualization, but they are in fact seldom realized due to the failure to adapt to the specifics of a corpus.[28] In bag-of-words representations, the surrogates into which text are abstracted are vectors that assign numerical values to each word in a list deemed relevant for the texts. Such quantification of literary texts naturally comes with the risk of arbitrariness and misinterpretation of artefacts produced by the approach itself.[29] A detailed understanding of every step in the process of operationalization is required to draw defensible conclusions. Much research on stylometric corpus analyses has focused on finding an adequate text distance measure as a means of authorship attribution.[30] In a deliberately compiled corpus, as opposed to a representatively sampled one, we would presuppose the existence of

ence extended abstracts, Vancouver, BC, Canada: ACM Press, 2011), 2269, https://doi.org/10.1145/1979742.1979906.

[27] Simon Päpcke et al., "Stylometric Similarity in Literary Corpora: Non-Authorship Clustering and 'Deutscher Novellenschatz'", *Digital Scholarship in the Humanities*, 38, no. 1 (2023): 277–95, 1–19, https://doi.org/10.1093/llc/fqac039.

[28] Even if there are interesting approaches, e.g. Laura Rettig, Regula Hanggli, and Philippe Cudre-Mauroux, "The Best of Both Worlds: Context-Powered Word Embedding Combinations for Longitudinal Text Analysis", in *2020 IEEE International Conference on Big Data (Big Data)* (2020 IEEE International Conference on Big Data, Atlanta, GA, USA: IEEE, 2020), 4741–50, https://doi.org/10.1109/BigData50022.2020.9377955.

[29] Andrew Piper, *Can We Be Wrong? The Problem of Textual Evidence in a Time of Data*, Cambridge Elements. Digital Literary Studies (Cambridge: Cambridge University Press, 2020), https://doi.org/10.1017/9781108922036.

[30] John Burrows, "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship", *Literary and Linguistic Computing* 17, no. 3 (1 September 2002): 267–87, https://doi.org/10.1093/llc/17.3.267; J. Rybicki, D. Hoover, and M. Kestemont, "Collaborative Authorship: Conrad,

meaningful groups of texts based on attributes other than authorship. Examples include genre, author gender, epoch, or narrative aspects; some of which may be reflected stylometrically. Indeed, even in the seemingly confined space of stylometric operationalizations of similarity, there are sufficient degrees of freedom to tailor them to the different kinds of groups.

Among the steps that lead from word frequencies to quantification of stylometric similarity are lemmatization, filtering, occurrences as well as dispersion, normalization, and the use of different aggregation functions. The resulting similarities are then interpreted through clustering and visualization.

Initial steps include the decision as to which lexical items are retained for further analysis. From a methodological point of view, the selection of terms can rule out undesired associations arising through artefacts from the chosen list of words. Here, the qualification as undesired should always be understood as relative to the underlying research question. Typical differences arising at this stage include decisions on case sensitivity, stemming, and lemmatization. As they are often considered explicitly,[31] we focus on two examples that are different in nature.

Table 1: First-person and other narrative perspectives.

relative frequency of ich vs. first-person narrative	first-person perspective	other perspective
above threshold	16	5
below threshold	5	60

Lists of frequent terms from the documents of a literary corpus usually contain a variety of personal and possessive pronouns. Thus, when we are applying stylometric similarities based on such a list, we can expect to find high degrees of such similarity between texts that are first-person narratives. Table 1 shows that there is a relative-frequency threshold for the first-person singular pronoun 'ich' that serves as a highly accurate classifier for first-person narratives. If, however, the narrative perspective

Ford and Rolling Delta", *Literary and Linguistic Computing* 29, no. 3 (1 September 2014): 422–31, https://doi.org/10.1093/llc/fqu016.

[31] E.g. Christopher D. Manning et al., "The Stanford CoreNLP Natural Language Processing Toolkit", in *Association for Computational Linguistics (ACL) System Demonstrations*, 2014, 55–60, http://www.aclweb.org/anthology/P/P14/P14-5010; Nils Reiter, Anette Frank, and Oliver Hellwig, "An NLP-Based Cross-Document Approach to Narrative Structure Discovery", *Literary and Linguistic Computing* 29, no. 4 (January 2014): 583–605, https://doi.org/10.1093/llc/fqu055.

is not the focus of the analysis, it seems to be more fruitful to replace pronouns in the list of frequent words by more generic terms.

Similar effects arise for names of places, countries, cities, or landscapes. While mentions of specific places are usually rare across an entire corpus, they often serve interchangeable roles in their respective novellas. Grouping them into generic tokens such as country, city, or countryside will alter the similarity of texts compared through these terms. A research question could then be guided by epoch theory and test, for example, the hypothesis that texts from Romanticism are placed in rural settings while those from German Realism have urban settings.[32]

The determination of the terms to be included in the analysis establishes the domain of frequency variables, but does not yet specify which values are to be assigned to them. Common choices include the raw counts of words and normalized values based on these counts. However, raw counts do not take into account the distribution of terms across a document. To be able to detect distributional differences, variables may be split into multiple counts related to the same term in different segments of the text. In the context of novellas, segmentation could, for example, be based on the common structuring of the stories within a story. In practice, this is done by comparing texts that contain words with similar document frequency: if we observe a strong accumulation of these terms in a few segments for some texts while others have an even distribution of the terms across the whole document, it suggests that the former texts indeed contain stories within a story.

The options named above are but examples of steps in the data preparation with a potential influence on how texts are related to one another. In order to actually relate the vectors of word frequencies, it is generally advisable to reduce their dimensionality. Otherwise, the word vectors would either have uneven lengths or a high number of zero elements, because many terms occur in only a few of the texts in a corpus. The effect of dimensionality of word vectors on text clustering was, for example, studied by Büttner.[33]

There are three typical filtering methods to avoid this phenomenon and they can also be applied independently. The first is dimensionality reduction by considering only a fixed number of the most frequent words. The second is the use of a stop word list containing common words (usually function words) that are present in each text and thus deemed irrelevant for the distinction of texts. A third possibility, referred to as culling, considers the number of documents in which a term appears: a term is considered only if it appears in a specified percentage of the texts in a corpus. This usually ensures that named entities are not part of the comparison; strong culling

[32] See Franco Moretti, *Graphs, Maps, Trees: Abstract Models for a Literary History* (London/New York: Verso, 2005).

[33] Andreas Büttner et al., "»Delta«in der stilometrischen Autorschaftsattribution", *Zeitschrift für digitale Geisteswissenschaften*, 2017, https://doi.org/10.17175/2017_006.

can, however, also lead to undesired side effects. In the *Novellenschatz*-corpus, we find that stylometric similarity discriminates the subgenre of Adelsnovellen if noble titles are part of the vector representation, but these titles are generally eliminated by the culling.

As a final example for the choice of representation on a distant reading scale, we consider the values assigned in word frequency vectors. Again there are three major options, the selections of which influence any subsequent analysis. A simple solution is the division of the raw frequency count by text length to make frequencies comparable in different documents. It yields the so-called term-frequency, or tf, score, which amplifies the influence of frequent words on the analysis compared to less frequent words. It can be modified by taking into account the number of documents in which a term appears, thus resulting in the term-frequency inverse document-frequency, or tf-idf, score.[34] This score favours words, which discriminate a text from the majority of the corpus, by attributing higher weights to them. A corresponding finding for the *Novellenschatz* is that novellas of female authors tend to cluster if term-frequency scores are used, owing to their greater use of female pronouns and articles. A scoring method, often used in authorship attribution tasks, is the standardization of term frequencies by subtracting the corpus mean and normalizing this difference by the standard deviation. This so-called z-score expresses higher- and lower-than-average frequencies in units of standard deviation.[35] For the *Novellenschatz*, z-scores prove useful to identify a group of novellas with an unusually high use of verbs in past tense and under-representation of words connected to direct speech. This hints at a group of texts with low prevalence of direct speech, which is deemed uncharacteristic for novellas, as the genre has been claimed to be similar to that of drama.[36]

We have thus argued that, even at a relatively fixed scale, any representation results from a long list of choices. The choices made have consequences and should therefore be made to align with the research question at hand. But it does not end there: when surrogates are processed further, the means of analysis need to be consistent with the goal as well, and not only on a principled level. In the present exam-

[34] See Lukáš Havrlant and Vladik Kreinovich, "A Simple Probabilistic Explanation of Term Frequency-Inverse Document Frequency (Tf-Idf) Heuristic (and Variations Motivated by This Explanation)", *International Journal of General Systems* 46, no. 1 (2017): 27–36, https://doi.org/10.1080/03081079.2017.1291635.

[35] Thomas Weitin, "Burrows's Delta Und Z-Score-Differenz Im Netzwerkvergleich. Analysen Zum Deutschen. Novellenschatz von Paul Heyse Und Hermann Kurz (1871–1876)", in Fotis Jannidis, Germanistische Symposien (ed.), *Digitale Literaturwissenschaft.* Beiträge Des DFG-Symposiums 2017 (Stuttgart: Metzler, 2023) (forthcoming).

[36] "die heutige Novelle ist die Schwester des Dramas", Theodor Storm, "Eine zurückgezogene Vorrede aus dem Jahre 1881", in Albert Köster (ed.), *Theodor Storms sämtliche Werke in acht Bänden*, vol. 8, 2012, 122–23.

ple, vectors of word frequencies, however constructed, are compared pairwise by similarity. Incidentally, the selection of an appropriate measure of similarity is no less important than the construction of the vectors themselves. In fact, the *Novellenschatz* contains examples of three novellas that can be ranked in any order in terms of pairwise similarities by choosing one of the three most common measures. This is mostly due to the different importance these measures assign to large deviations in the scores of single entries.

Conclusion

The reading-at-scale approach critically reflects the trade-off that arises in computational literary analysis: context is reduced to enforce comparability, or comparability is given up for contextualization. Different reduction techniques incorporate a variety of scales for comparison, allowing to view the object of study from different perspectives.

In this chapter, we have demonstrated how different text characteristics affect the values obtained from text distance measures: as the latter are most often computed by averaging distances between words, several signals may be moving and transforming through stages of the analysis. The standard steps for the preprocessing of literary texts for digital analyses, such as tokenization,[37] aim at making texts more comparable and do not focus on eliminating or controlling specific signals. To be able to do that, operationalizations must be reflected upon meticulously and chosen in awareness of potential outcomes that would be considered as artefacts of a method.

With careful consideration, the detection of signals relevant to literary scholarship may be supported by the appropriate choice of representations and parametrization of methods. Beyond existing simplistic suggestions that certain most frequent words harbour the authorship signal, while less frequent words contain the signal for a literary epoch, we have found a variety of other more complex signals in the *Novellenschatz*-corpus. These signals can be differentiated through additional preparatory steps. For instance, a similar narrative perspective and similar plot elements account for smaller distances between texts along with authorship, epoch, and sometimes even the protagonist's social class.

Further, we observed that pronouns were strongly associated with the narrator's perspective. Therefore, if this perspective is not relevant for the main research question, it may be necessary to undertake alterations to the text surrogate to eliminate this signal. Similarly, toponyms are candidates for alterations: if a project aims to

[37] E.g. Matthew L. Jockers and Rosamond Thalken, *Text Analysis with R for Students of Literature*, 2nd ed. (Springer, 2020), https://doi.org/10.1007/978-3-030-39643-5.

examine the hierarchy between geographical places in rural settings, the toponym could be changed to a broader type of place.

It is important to keep in mind that signals discovered so far are specific to the *Deutscher Novellenschatz*-corpus. In subsequent research, detailed analyses of the other collections are planned. This order of exploration is motivated by the historical dimension of the collections: the *Deutscher Novellenschatz* is considered a genre-defining collection in literary historiography, and we expect our findings about the novella genre to be solidified on the basis of results related to the *Neuer Deutscher Novellenschatz*, thus further exploring the question whether the *Deutscher Novellenschatz*-collection can be considered a representative corpus for novellas of the 19th century.[38] As this article demonstrates, reading at scale links digital methods and operationalizations to the middle-sized *Novellenschatz*-corpus as a historical artefact, gradually including insights from data analysis and research in literary history in the adaptations of operationalizations. The concept allows switching between abstract representations of literary texts (such as word vectors and lists of most frequent words) and analytical text interpretations in the hermeneutic tradition, thus bringing both into a fruitful exchange.

Bibliography

Burrows, John. "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship". *Literary and Linguistic Computing* 17, no. 3 (1 September 2002): 267–87. https://doi.org/10.1093/llc/17.3.267.

Büttner, Andreas, Friedrich Michael Dimpel, Stefan Evert, Fotis Jannidis, Steffen Pielström, Thomas Proisl, Isabella Reger, Christof Schöch, and Thorsten Vitt. "»Delta«in der stilometrischen Autorschaftsattribution". *Zeitschrift für digitale Geisteswissenschaften*, 2017. https://doi.org/10.17175/2017_006.

Erlin, Matt, and Lynne Tatlock. "Introduction: 'Distant Reading' and the Historiography of Nineteenth Century German Literature". In *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, edited by Matt Erlin and Lynne Tatlock, 1–29. Studies in German Literature, Linguistics, and Culture. Rochester, New York: Camden House, 2014.

Gervinus, Georg Gottfried. *Handbuch Der Geschichte Der Poetischen National-Literatur Der Deutschen*. 3. Aufl. Leipzig: Engelmann, 1844.

Ghannay, Sahar, Benoit Favre, Yannick Estève, and Nathalie Camelin. "Word Embedding Evaluation and Combination". In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 300–305. Portorož, Slovenia:

[38] Fotis Jannidis, "Perspektiven quantitativer Untersuchungen des Novellenschatzes", *Zeitschrift für Literaturwissenschaft und Linguistik* 47, no. 1 (March 2017): 7–27, https://doi.org/10.1007/s41244-017-0050-x.

European Language Resources Association (ELRA), 2016. https://aclanthology.org/L16-1046.

Gius, Evelyn, Jan Christoph Meister, Malte Meister, Marco Petris, Christian Bruck, Janina Jacke, Mareike Schumacher, Dominik Gerstorfer, Marie Flüh, and Jan Horstmann. "CATMA 6", 6 April 2022. https://doi.org/10.5281/ZENODO.1470118.

Havrlant, Lukáš, and Vladik Kreinovich. "A Simple Probabilistic Explanation of Term Frequency-Inverse Document Frequency (Tf-Idf) Heuristic (and Variations Motivated by This Explanation)". *International Journal of General Systems* 46, no. 1 (2017): 27–36. https://doi.org/10.1080/03081079.2017.1291635.

Heintz, Bettina. "Numerische Differenz. Überlegungen zu einer Soziologie des (quantitativen) Vergleichs / Numerical Difference. Toward a Sociology of (Quantitative) Comparisons". *Zeitschrift für Soziologie* 39, no. 3 (1 January 2010). https://doi.org/10.1515/zfsoz-2010-0301.

Hengchen, Simon, Ruben Ros, Jani Marjanen, and Mikko Tolonen. "A Data-Driven Approach to Studying Changing Vocabularies in Historical Newspaper Collections". *Digital Scholarship in the Humanities* 36, no. Supplement 2 (5 November 2021): 109–26. https://doi.org/10.1093/llc/fqab032.

Heyse, Paul, and Hermann Kurz. "Einleitung". In *Deutscher Novellenschatz*, Vol. 1. München: R. Oldenbourg, 1871. https://www.deutschestextarchiv.de/heysekurz_einleitung_1871.

Horstmann, Jan. "Undogmatic Literary Annotation with CATMA. Functions, Differentiation, Systematization". In *Annotations in Scholarly Editions and Research*, edited by Julia Nantke and Frederik Schlupkothen, 157–76. De Gruyter, 2020. https://doi.org/10.1515/9783110689112-008.

Jannidis, Fotis. "Perspektiven quantitativer Untersuchungen des Novellenschatzes". *Zeitschrift für Literaturwissenschaft und Linguistik* 47, no. 1 (March 2017): 7–27. https://doi.org/10.1007/s41244-017-0050-x.

Jockers, Matthew L. *Macroanalysis: Digital Methods and Literary History*. Topics in the Digital Humanities. Urbana, IL: University of Illinois Press, 2013.

Jockers, Matthew L., and Rosamond Thalken. *Text Analysis with R for Students of Literature*. 2nd ed. Springer, 2020. https://doi.org/10.1007/978-3-030-39643-5.

Kleymann, Rabea. "Datendiffraktion: Von Mixed zu Entangled Methods in den Digital Humanities". HTML,XML,PDF. Edited by Manuel Burghardt, Lisa Dieckmann, Timo Steyer, Peer Trilcke, Niels Walkowski, Joëlle Weis, and Ulrike Wuttke. *Fabrikation von Erkenntnis – Experimente in den Digital Humanities (Zeitschrift für digitale Geisteswissenschaften / Sonderband)*, 2022. https://doi.org/10.17175/SB005_008.

Krautter, Benjamin, Janis Pagel, Nils Reiter, and Marcus Willand. "Eponymous Heroes and Protagonists – Character Classification in German-Language Dramas". *LitLab Pamphlet* #7 (2018): 58.

Lorenzini, Matteo, Marco Rospocher, and Sara Tonelli. "On Assessing Metadata Completeness in Digital Cultural Heritage Repositories". *Digital Scholarship in the Humanities* 36, no. Supplement 2 (5 November 2021): 182–88. https://doi.org/10.1093/llc/fqab036.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. "The Stanford CoreNLP Natural Language Processing Toolkit". In *Association for Computational Linguistics (ACL) System Demonstrations*, 55–60, 2014. http://www.aclweb.org/anthology/P/P14/P14-5010.

Moretti, Franco. "Conjectures on World Literature". *New Left Review* 1 (2000): 54–68.

Moretti, Franco. "Conjectures on World Literature". In *Distant Reading*, 43–62. London; New York: Verso, 2013.

Moretti, Franco. *Graphs, Maps, Trees: Abstract Models for a Literary History*. London; New York: Verso, 2005.

Mueller, Martin. "Morgensterns̒ Spectacles or the Importance of Not-Reading". *Northwestern University Digital Humanities Laboratory* (blog), 2013. https://sites.northwestern.edu/nudhl/?p=433.

Mueller, Martin. *Scalable Reading*, 2019. https://scalablereading.northwestern.edu/.

Päpcke, Simon, Thomas Weitin, Katharina Herget, Anastasia Glawion, and Ulrik Brandes. "Stylometric Similarity in Literary Corpora: Non-Authorship Clustering and 'Deutscher Novellenschatz'". *Digital Scholarship in the Humanities*, 38, no. 1 (2023): 277–95, 1–19. https://doi.org/10.1093/llc/fqac039.

Piper, Andrew. *Can We Be Wrong? The Problem of Textual Evidence in a Time of Data*. Cambridge Elements. Digital Literary Studies. Cambridge: Cambridge University Press, 2020. https://doi.org/10.1017/9781108922036.

Piper, Andrew. *Enumerations: Data and Literary Study*. Chicago, IL; London: The University of Chicago Press, 2018.

Reiter, Nils, Anette Frank, and Oliver Hellwig. "An NLP-Based Cross-Document Approach to Narrative Structure Discovery". *Literary and Linguistic Computing* 29, no. 4 (January 2014): 583–605. https://doi.org/10.1093/llc/fqu055.

Rettig, Laura, Regula Hanggli, and Philippe Cudre-Mauroux. "The Best of Both Worlds: Context-Powered Word Embedding Combinations for Longitudinal Text Analysis". In *2020 IEEE International Conference on Big Data (Big Data)*, 4741–50. Atlanta, GA, USA: IEEE, 2020. https://doi.org/10.1109/BigData50022.2020.9377955.

Rybicki, J., D. Hoover, and M. Kestemont. "Collaborative Authorship: Conrad, Ford and Rolling Delta". *Literary and Linguistic Computing* 29, no. 3 (1 September 2014): 422–31. https://doi.org/10.1093/llc/fqu016.

Scherer, Wilhelm. *Geschichte Der Deutschen Litteratur*. 3. Auflage. Berlin: Weidmannsche Buchhandlung, 1885.

Scherer, Wilhelm, and Oskar Walzel. *Geschichte Der Deutschen Literatur.* Berlin: Askanischer Verlag, 1921.

Schöch, Christof. “Big? Smart? Clean? Messy? Data in the Humanities”. *Journal of Digital Humanities* 2 (2013): 1–12.

Storm, Theodor. “Eine zurückgezogene Vorrede aus dem Jahre 1881”. In *Theodor Storms sämtliche Werke in acht Bänden*, edited by Albert Köster, 8:122–23, 2012.

Thai, VinhTuan, and Siegfried Handschuh. “Context Stamp: A Topic-Based Content Abstraction for Visual Concordance Analysis”. In *Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems – CHI EA ’11*, 2269. Vancouver, BC, Canada: ACM Press, 2011. https://doi.org/10.1145/1979742.1979906.

Walkhoff, Monika. *Der Briefwechsel Zwischen Paul Heyse Und Hermann Kurz in Den Jahren 1869 – 1873 Aus Anlass Der Herausgabe Des Deutschen Novellenschatzes [Mit Faks.]*. München: Foto-Druck Frank, 1967.

Walzel, Oskar. *Die Deutsche Dichtung Seit Goethes Tod*. Berlin: Askanischer Verlag, 1919.

Weitin, Thomas. “Average and Distinction. The Deutsche Novellenschatz Between Literary History and Corpus Analysis”, LitLab Pamphlet, 6 (2018): 1–23.

Weitin, Thomas. “Burrows‘s Delta Und Z-Score-Differenz Im Netzwerkvergleich. Analysen Zum Deutschen. Novellenschatz von Paul Heyse Und Hermann Kurz (1871–1876)”. In *Digitale Literaturwissenschaft. Beiträge Des DFG-Symposiums 2017*, edited by Fotis Jannidis. Germanistische Symposien. Stuttgart: Metzler, 2023.

Weitin, Thomas. *Digitale Literaturgeschichte. Eine Versuchsreihe in 7 Experimenten*. Berlin: Metzler / Springer Nature, 2021.

Weitin, Thomas. “Thinking Slowly. Reading Literature in the Aftermath of Big Data”. *LitLab Pamphlet* #1 (2015): 1–19.

Weitin, Thomas, and Katharina Herget. “Falkentopics: Über Einige Probleme Beim Topic Modeling Literarischer Texte”. *Zeitschrift Für Literaturwissenschaft Und Linguistik* 47, no. 1 (2017): 29–48. https://doi.org/10.1007/s41244-017-0049-3.

Corpora

Weitin, Thomas. Digitalized text corpus. ‘Der Deutsche Novellenschatz’. Edited by Paul Heyse, Hermann Kurz. 24 volumes, 1871–1876. Darmstadt/Konstanz, 2016. https://www.deutschestextarchiv.de/doku/textquellen#novellenschatz.

Weitin, Thomas and Herget, Katharina. Digitalized text corpus. ‘Der Neue Deutsche Novellenschatz’. Edited by Paul Heyse, Ludwig Laistner. 24 volumes, 1884–1887. Darmstadt, 2022. DOI: 10.5281/zenodo.6783577.

Weitin, Thomas and Herget, Katharina. Digitalized text corpus. ‘Novellenschatz des Auslandes’. Edited by Paul Heyse, Hermann Kurz. 14 volumes, 1877–1884. Darmstadt, 2022. DOI: 10.5281/zenodo.6784080.