



Modeling and Predicting Literary Reception A Data-Rich Approach to Literary Historical Reception

Judith Brottrager¹ 
Annina Stahl² 
Arda Arslan²
Ulrik Brandes² 
Thomas Weitin¹ 

1. LitLab, Technical University Darmstadt , Darmstadt, Germany.
2. Social Networks Lab, ETH Zurich , Zurich, Switzerland.

Citation

Judith Brottrager, Annina Stahl, Arda Arslan, Ulrik Brandes, and Thomas Weitin (2022). "Modeling and Predicting Literary Reception. A Data-Rich Approach to Literary Historical Reception". In: *Journal of Computational Literary Studies* 1 (1). [10.48694/jcls.95](https://doi.org/10.48694/jcls.95)

Date published 2022-11-24

Date accepted 2022-03-30

Date received 2021-12-22

Keywords

historical reception, operationalization, sentiment analysis, text classification, 18th century, 19th century

License

CC BY 4.0 

Reviewers

Katherine Bode 

Karina van Dalen-Oskam 

Note

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

Abstract. This contribution exemplifies a workflow for the quantitative operationalization and analysis of historical literary reception. We will show how to encode literary historical information in a dataset that is suitable for quantitative analysis and present a nuanced and theory-based perspective on automated sentiment detection in historical literary reviews. Applying our method to corpora of English and German novels and narratives published from 1688 to 1914 and corresponding reviews and circulating library catalogs, we investigate if a text's popularity with lay audiences, the attention from contemporary experts or the sentiment in experts' reviews can be predicted from textual features, with the aim of contributing to the understanding of how literary reception as a social process can be linked to textual qualities.

1. Introduction

For traditional literary studies approaches, the text itself is hardly ever the only subject of investigation when addressing questions related to developments in literary history. Instead, a wide range of complementary data, from letters to reviews and poetological treatises, are employed to embed a text, its production, and its reception in a broader literary historical context. Such a richness of detail and context is *per definitionem* not achievable when working with quantitative methods: When analyzing hundreds or thousands of texts, linking each and every one of them to their immediate context of production and reception is simply not feasible. The first hurdle of such a context-heavy quantitative approach is the lack of available data. In comparison to the entire mass of literary history, there are only few literary works which have been researched thoroughly enough to be described on all levels of production and reception. The second hurdle is that of formalization and operationalization. Even if qualitative research about all texts was available, this unstructured data would need to be digitized and operationalized to be used for quantitative analysis, again leading to a loss of detail.

Building on context-sensitive approaches suggested in previous research, it is the aim of this contribution to find an appropriate level of abstraction in "data-rich literary history" (Bode 2018, 37–57) by exemplifying a workflow for the quantitative operationalization and analysis of historical literary reception, and to use this newly formalized data

to investigate if external markers of reception can be predicted from features of the texts themselves. In the course of this paper, we will (1) show how to encode literary historical information in a dataset that is suitable for quantitative analysis, and apply this method to a collection of roughly 1,200 English and German novels and narratives published between 1688 and 1914 along with data on the reception of these works by their contemporaries, (2) present a nuanced and theory-based perspective on automated sentiment detection in historical literary reviews, and (3) compare contemporary experts' reviews and a text's popularity to textual features that reflect a text's complexity and distinctiveness.

As part of a greater research interest in the comparative analysis of canonization processes in English and German literary history (see Brottrager et al. 2021), our approach operates between the poles of a text's canonization status today – a result of a myriad of stacked selection processes – and its reception by its immediate contemporaries. The comparison between English and German literary history seems especially fruitful here, as their classical periods are temporally and philosophically far apart. The German classical period from 1770 to 1830 with its focus on the authorial genius and aesthetic autonomy remains a figurative yardstick for subsequent generations of writers and critics, ingraining the dichotomy of high and low literature in German literary history (Heydebrand and Winko 1996, 151–157), while such a stark distinction is not encoded in English literary history. By comparing these two very different traditions over a time span that encompasses the German Classicism, but also the rise of the novel and the so-called 'Novellenflut' as phenomena of popular fiction, we will be able to show how initially well-received literary texts get lost in the so-called 'Great Unread', while others are elevated into the canon.

We will begin by discussing examples of context-rich approaches to literary reception and previous research on the categorization of reviews in the context of computational literary analyses (Section 2). This overview of practical applications will then be followed by an in-depth examination of the theoretical background of verbal judgments and evaluative actions in literary reception. Following the description of our canon-conscious corpus selection, the paper's third and fourth section will show how historical sources of literary information can be encoded in a dataset by adding reviews as representations of verbal value judgments (Section 3) and circulating library catalogs as proxies for audiences' interests (Section 4). The methodological part of this contribution (Section 5) will show how we have implemented a SentiArt-inspired approach (A. M. Jacobs 2019) to evaluative language for the differentiation of literary reviews. Then, we present how we used the historical data introduced in previous sections to analyze to which extent the popularity and reception of literary works can be explained with qualities of the texts themselves (Section 6). In the discussion (Section 7), we will illustrate how the theoretical framework of historical evaluation is reflected in our results.

2. Previous Work

While the examination of text-related metadata categories, such as authorial gender, genre, publication date, and broad thematic categories has already been introduced in early contributions to the field of Computational Literary Studies (CLS) (Jockers

2013; Moretti 2013), the study of reception-related data is not yet as established. Some studies have suggested measures of prestige and popularity (Algee-Hewitt et al. 2016; Porter 2018; Underwood 2019; Underwood and Sellers 2016), where these categories reflect to some degree reception-related aspects: In their publication on literary prestige, Underwood and Sellers define prestige as a dichotomy by distinguishing poems according to whether or not they were reviewed in prestigious journals (2016, 323–325, see also Underwood 2019, 68–110). Algee-Hewitt et al. (2016) similarly determine a text's prestige, but do so by operationalizing the category as the number of bibliographical entries in the MLA featuring the author as the "Primary Subject Author". Additionally, they introduce the category of popularity, modeled as the combination of the number of reprints and translations (2016, 3). Capturing modern readers' responses, Porter (2018) constructs a score representing the popularity of authors by combining Goodreads metrics (the number of ratings, the number of reviews, and the author's average rating).

The hesitation to include historical reviews as actual textual data seen in the examples above is understandable: Reviews often have to be retro-digitized before they can be analyzed, and established methods developed for categorizing shorter, more straightforward modern language reviews such as sentiment analysis are not as reliable when confronted with historical language. Du and Mellmann (2019) address these issues and suggest a multi-layered approach when dealing with historical reviews: Instead of relying solely on lexicon-based sentiment analysis,¹ they aggregated a metric that takes the distance between sentiment expression and author name into account to ensure that value judgments directly connected to an author's work are more strongly weighted. Combined with textual features such as (lemmatized) *n*-grams with weights based on tf-idf and word embeddings, these sentiment values were then used to train a Support Vector Machine (SVM) which correctly identified positive, negative, and neutral sentences extracted from reviews with an overall average accuracy of 0.64 and up to 0.76 for only positive and negative sentences (Du and Mellmann 2019, 11).

When discussing the historical specificity of literary reviews and their implicitly marked registers (2019, 13), Du and Mellmann hint at elements of verbal judgments that are also extensively investigated by Heydebrand and Winko (1996) in their introductory work on evaluation in literature. According to Heydebrand and Winko (1996, 62), verbal value judgments can be defined as illocutionary acts of utterance through which an object is ascribed an attributive value. This attributive value in turn links back to a defined value system.² Different value systems lead to different attributive values: While in one historical context a specific characteristic is seen as valuable, it can be ascribed less value in another historical period (Heydebrand and Winko 1996, 111–131, 134–162).

In addition to verbal value judgments, Heydebrand and Winko elaborate on social components of evaluation, especially those connected to selection processes. They point out that decisions for or against a text are evaluative operations that structure

1. Du and Mellmann use a manually modified version of the German sentiment lexicon SentiWS (Remus et al. 2010).

2. Heydebrand and Winko 1996, 62: "Sprachliche Wertungen zählen zu den illokutionären Akten. Von anderen Sprechhandlungen unterscheiden sie sich durch eine besondere Art der Zuschreibungsbeziehung: Sie schreiben einem Objekt mittels eines Wertausdrucks einen attributiven Wert zu, und zwar auf der Grundlage eines axiologischen Werts und bestimmter Zuordnungsvoraussetzungen." ("Verbal value judgements belong to the illocutionary acts. They differ from other speech acts by a special kind of attributive relation: They attribute a value to an object by means of a value expression, on the basis of an axiological value and certain attributive presuppositions." Translation by the authors).

all levels of the literary system, from a publisher's acceptance of a manuscript to a reader's individual buying decision (1996, 79). Selective decisions by literary critics³ are especially impactful, as the existence of professional reviews spotlights a text when compared to the mass of all other published but unreviewed competitors (1996, 99).

Similar to our previous work on the issue of canonization (Brottrager et al. 2021), introducing an operationalization for contemporary reception based on the theoretical framework provided by Heydebrand and Winko (1996) aims at creating comparability within our own project, but is also part of a greater effort in the field of CLS to find suitable, reproducible, and adaptable implementations for complex literary concepts (see Alvarado 2019; Pichler and Reiter 2021; Schröter et al. 2021).

3. Corpora

For the compilation of our two corpora, we systematically adapted an approach proposed by Algee-Hewitt and McGurl (2015) in their contribution on creating a balanced novel corpus for the 20th century. To tackle what they call "dilemmas of selection" (2015, 1), they combine existing best-of and bestseller lists with commissioned lists of novels suggested by experts of Feminist and Postcolonial Studies to create a corpus that entails multiple dimensions of canonicity: First, a very narrowly defined normative canon of the 'best' novels written in the 20th century, second, financially successful and thus presumably popular novels, and third, novels belonging to an alternative canon of marginalized texts. In contrast to "samples of convenience" usually found in readily available online collections, which are "no doubt equally, if not more biased than the lists we have assembled" (2015, 22), using a predefined corpus list allows for a monitoring of availability issues and canonical biases.

For corpora covering the Long 18th and 19th Century (1688-1914), comparable lists are not or only partially available. To be able to still apply a similar logic, we had to find a way to adequately replace both existing and commissioned lists. As described above, the lists represent different dimensions of the canon, which can also be replicated when using lists of mentions extracted from differently motivated literary histories and other secondary sources. By relying on lists of texts deemed relevant by experts with different focal points, we would still be able to contrast the "found" corpus (2015, 4) of already digitized material with a "made" list (2015, 15) of, if not commissioned, but still purposely gathered texts. To capture the essence of normative best-of lists, we used highly condensed and consequently exclusive narrative literary histories and anthologies. Lists of popular literature and marginalized literary texts were reconstructed by including specialized sources (e.g. sources on light fiction and popular genres, companions to literature by female authors and literature from geographical peripheries) and by surveying the broader academic canon (e.g. companions to specific genres and periods).

The resulting list was then used as a basis for the corpus compilation. In a first iteration, we checked online full-text repositories.⁴ For texts not available as digitized full-text,

3. Heydebrand and Winko call them and other professional agents in the literary field "Verarbeiter" (= processors) (1996, 99)

4. Textgrid, Deutsches Textarchiv (DTA), Eighteenth Century Collections Online (ECCO), Project Gutenberg US, Projekt Gutenberg-DE, Project Gutenberg Australia, Project Gutenberg Canada, Sophie, ebooks@Adelaide (no longer available, but still accessible through the Internet Archive).

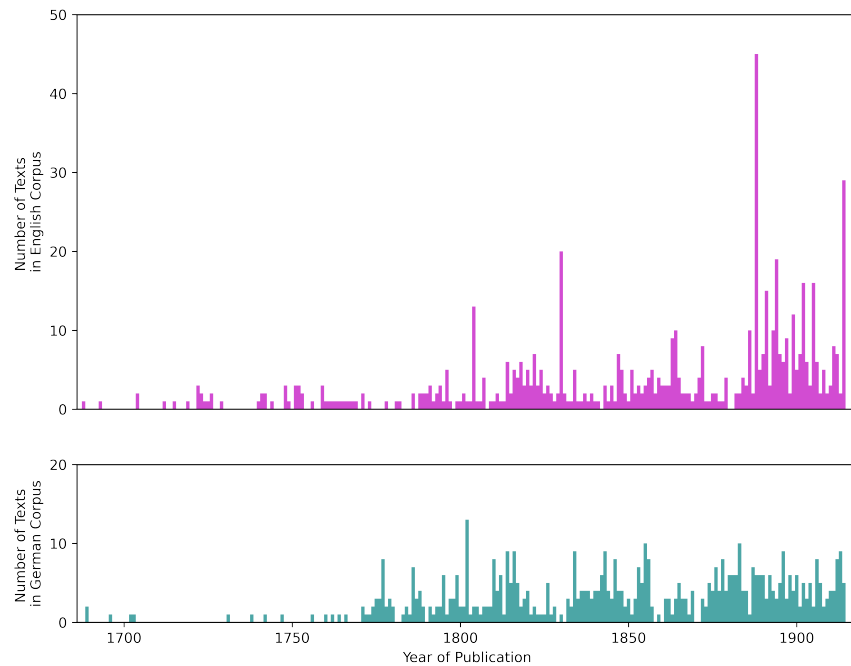


Figure 1: Temporal distribution of texts in our corpora.

we looked for high-quality scans or scanned and retro-digitized them ourselves. As Algee-Hewitt et al. (2016, 2) point out, the retro-digitization is cost- and time-intensive, which is why we did not retro-digitize all missing entries, but deliberately included texts that added a degree of diversity to our corpus because they were written by an author not already included, represent a niche genre, or other forms of marginalized literature. To ensure high-quality transcriptions, the workflow combines automated optical character recognition (OCR) and manual post-corrections.

The compilation resulted in an English corpus of 605 and a German corpus of 547 texts. The temporal distribution of publication dates in both corpora is shown in Figure 1. In both corpora, the number of texts increases around 1770, which corresponds to historically informed expectations linked to the rise of the novel in both English and German literary history. Later spikes in the English corpus are primarily caused by the inclusion of collections of (short) stories, which are incorporated as individual texts.

4. Complementary Data

To be able to model literary evaluation as described by Heydebrand and Winko (1996), we expanded our dataset to include representations of verbal value judgments and readers' selective choices. While verbal value judgments are directly preserved in historical reviews, the reconstruction of readers' choices is not as straightforward. Transferring Heydebrand and Winko's idea of the buying decision to the time frame in question seems impractical because particularly for earlier time periods covered by our corpora, reliable sales numbers are not available. Additionally, we wanted to introduce a measure

that explicitly encapsulates a text's popularity with lay audiences in contrast to expert opinions recorded in reviews, and historically, buying books was simply not the way the majority of readers accessed their reading materials. Here, entries in circulating library catalogs seem to be a better suited proxy: Circulating libraries relied heavily on the popularity of the items they advertised and had to adapt to audiences' preferences in order to remain profitable (E. H. Jacobs 2003), which makes the existence of catalog entries a suitable representation of a text's popularity.

4.1 Reviews

In both the English and German-speaking Europe, the rise of literary periodicals coincides with the commercialization of the literary market (see Italia 2012), which led to an exponential growth of available reading material and a resulting need for selection. As a consequence, literary periodicals can be seen as structuring devices (Plachta 2019) that place the reviewed texts along a gradient from well to poorly received, but also distinguish between texts that were interesting enough to be reviewed and the remaining mass of texts published at the same time. In addition to reviews being written by professional readers, numerous influential publications were directly linked to central figures of the literary sphere: Authors such as August Friedrich Kotzebue and Tobias Smollett, for example, acted as founders and editors of the *Blätter für literarische Unterhaltung* and *The Critical Review*, respectively. This direct involvement of authors as professional reviewers (see Heydebrand and Winko 1996, 188–210) further accentuate the difference between evaluations by (peer) experts and popularity with broader audiences, as it is recorded in circulating library catalogs described below.

Due to the sheer number of literary journals published in the time span covered by our corpora, the selection of representative journals is based on considerations of influence and outreach, but also availability. For the English dataset, we were able to rely on some already digitized reviews accessible through the database *British Fiction 1800-1829* (Garside 2011, based on Garside and Schöwerling 2000) and used the corresponding analogue bibliography for the time span from 1770-1799 (Raven and Forster 2000) to locate referenced reviews. The database and bibliography primarily list reviews in *The Monthly Review* (MR) (covering the years from 1800 to 1830) and *The Critical Review* (CR) (1800-1817), but also feature references to *La Belle Assemblée* (BA) (1806-1830), *Flowers of Literature* (FL) (1801-1809), and *The Star* (surveyed for 1800 through 1830). Additionally, we consulted the database *The Athenaeum Project* (ATH) (City University London 2001) which provides access to searchable indices of the eponymous journal published from 1828 to 1923. For the German dataset, we consulted the database *Gelehrte Journale und Zeitungen der Aufklärung* (GJZ18 2021), but also relied heavily on the monthly and yearly indices of selected journals which were especially influential during their respective running time: *Allgemeine Literatur-Zeitung* (ALZ) (1785-1849), *Morgenblatt für gebildete Stände* (MGS) (1807-1865), *Blätter für literarische Unterhaltung* (BLU) (1826-1898), and *Deutsche Literaturzeitung* (DL) (1880-1993).

As the available scan quality as well as the fonts and type settings differed widely across the selected publications, we trained multiple recognition models using OCR4all (Reul et al. 2019), which were then combined in several iterations of text recognition. Collective reviews of multiple texts were split into parts concerning the referenced

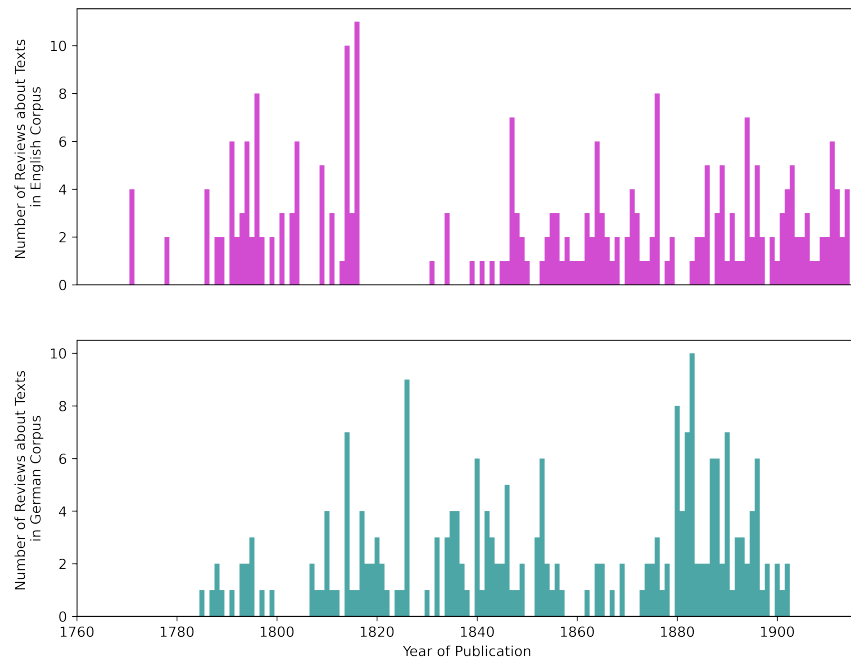


Figure 2: Temporal distribution of reviewed texts.

texts, and frequently featured lengthy quotes from the reviewed texts were replaced by ellipses.

In sum, we have collected 254 English and 221 German reviews. As some of them address the same texts, this results in 197 reviewed texts in the English and 176 reviewed texts in the German corpus, which means that we were able to link almost a third of each corpus to at least one historical review. Figure 2 shows the temporal distribution of reviews for both corpora. With the exception of a major gap in reviews concerning English texts from 1820 to 1830, which is most likely caused by the running time of the surveyed journals, the reviews are quite evenly distributed from 1780 onward. The lack of data before 1780 can again be linked to the selected journals, which is why all textual analyses (see Section 6) will take this bias into account.

4.2 Circulating Libraries

Similar to the emergence of literary journals, the introduction of circulating libraries is closely associated with the explosion of publication numbers related to the rise of the novel and the revolution of reading in the second half of the 18th century (Martino 1990, 1–134). By lending books to people who, as Gamer puts it, “would never have considered buying fiction” (2000, 65), circulating libraries can be seen as a form of democratizing literary consumption. However, the libraries’ broadening target group also caused concern with contemporaries, who warned against the moral corruption caused by circulating libraries’ focus on crowd-pleasing light literature (Jäger 1982, 263–264). Despite this criticism, circulating libraries became essential actors in the 19th century literary market, with some libraries, such as *Mudie’s Circulating Library*, gaining

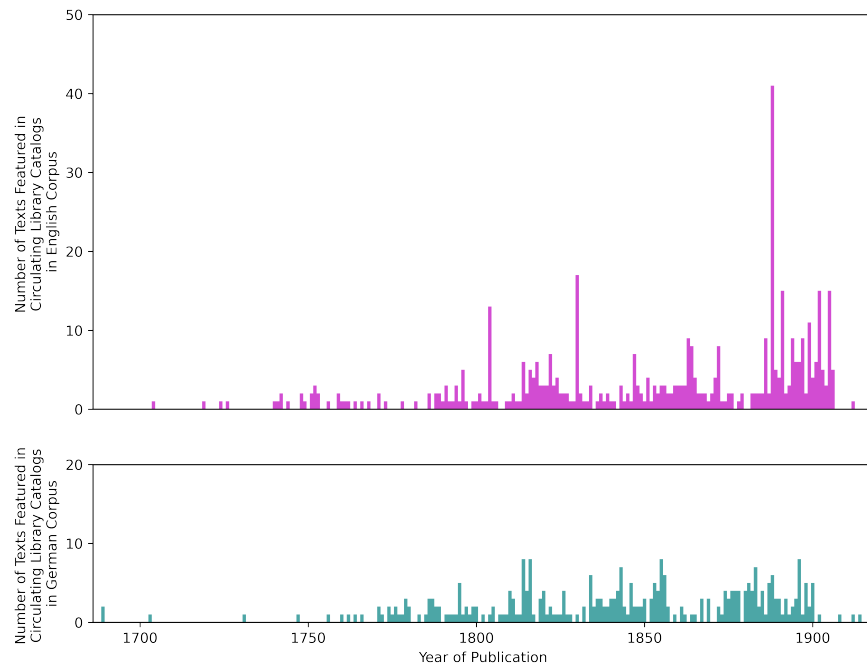


Figure 3: Temporal distribution of entries in circulating library catalogs.

so much influence and “purchasing power” that they could single-handedly “sell or condemn a book” (Katz 2017, 405).

Analogously to our approach to literary journals, the selection of specific catalogs is determined by questions of importance and availability. The issue of availability is more salient in this case: Compared to the number of preserved and recorded catalogs (Martino 1990, 917–1017), only very few of them are available as digital surrogates, which limits our options quite significantly. Nonetheless, we managed to find four English and six German catalogs published between 1809 and 1907 and 1790 and 1901, respectively, allowing for an adequate coverage of the 19th century. For the English dataset, we surveyed the 1809 catalog of *W. Storry’s General Circulating Library* (York), the 1829 catalog of *Hookham’s Library* (London), and two catalogs (1873 and 1907) for *Mudie’s Select Library* (London). Due to the municipal library of Vienna’s digital research focus on library catalogs, the German dataset is heavily skewed towards Viennese libraries and includes the 1790 and 1812 catalogs of rentable books at *Johann Georg Binz’* bookstore, *Carl Armbruster’s* 1813 catalog, *J. August Bachmann’s* 1851 catalog, *Friedrich Gerold’s* 1850 catalog, and the 1901 catalog of the *Literatur-Institut Ludwig und Albert Last*. Linking our corpus texts with entries in these catalogs required a two-step approach: Due to the diverging formats and indexing methods and inconsistent titles and spelling variations, we combined a full-text search of automatically recognized text with a manual double-check of indices for each catalog.

Of all 1,153 novels and narratives in our corpora, 763 were referenced in at least one of the catalogs we surveyed. Especially the coverage for the English corpus is significant: 75.54 percent of all texts and 78.57 percent of all featured authors appear in at least

one catalog. The same is true for 55.94 percent of all German texts and 54.34 percent of all German-speaking authors. The temporal distribution of texts available in library catalogs is presented in Figure 3. Whereas the circulating library entries for the German texts are quite evenly distributed from 1780 to 1914, there is more variance in the English corpus. From 1780 to 1890, the mean number of texts referenced in a catalog per year is 3.37, while for the years after 1890, the mean rises to 7.96. This is certainly due to the inclusion of collections of stories mentioned in Section 3, but also indicates that the last English catalog published in 1907 features many recent publications.

5. Methods

With our text collections and complementary historical reception data being made available for quantitative analysis, we investigated whether a text's reception can be linked to certain textual qualities. For this, we formalized and summarized reviews with sentiment analysis. We employed both an established and a custom sentiment analysis tool and assigned a sentiment score to each review. Then, we extracted textual features from our corpus texts that represent a text's lexical and syntactic complexity and its distinctiveness within the corpus. Based on these features and the reception data, we trained classifiers to predict the popularity with both reviewers and lay audience and a regression model to predict the sentiment scores of reviews.

5.1 Evaluative Language in Reviews

As described above, a basic sentiment analysis alone often fails to detect differences between historical reviews (Du and Mellmann 2019). This is partially due to the tools being designed for modern language usage but also due to specificities of evaluative language in literary reviews. When examining the collected reviews, it becomes apparent that especially negative reviews are often vague in their criticism and balance out criticism by mentioning minor positive aspects. Additionally, the reviews differ significantly in length – some of them consist of only a few sentences, while others span over several pages, featuring detailed plot synopses. Unsurprisingly, tools such as TextBlob (Loria 2018) and its extension for German, textblob-de (Killer 2019), are often not able to detect these subtleties. In a preliminary experiment with a test set of 15 positive and negative reviews for each dataset, TextBlob correctly identified all 15 positive English reviews and 13 positive German reviews, but only 8 negative English and 6 negative German reviews. With precision rates of 68.18 and 59.09 percent, we decided to implement an alternative approach using word embeddings to define the positive and negative poles of evaluative language in the specific context of historical reviews.

From a linguistic point of view, the evaluative language to be detected is an instance of appraisal (Halliday and Matthiessen 2014, Martin and White 2007). To be able to include not only explicit evaluative expressions on the word-level (e.g. “this is an excellent novel”) but also more implicit forms of appraisal (e.g. the positive connotation of “Gestalt” and negative connotation of “Geschöpf” described by Du and Mellmann 2019, 13) we ascribe words a value that represents their similarity to explicit evaluative expressions by calculating their distances in word embeddings.

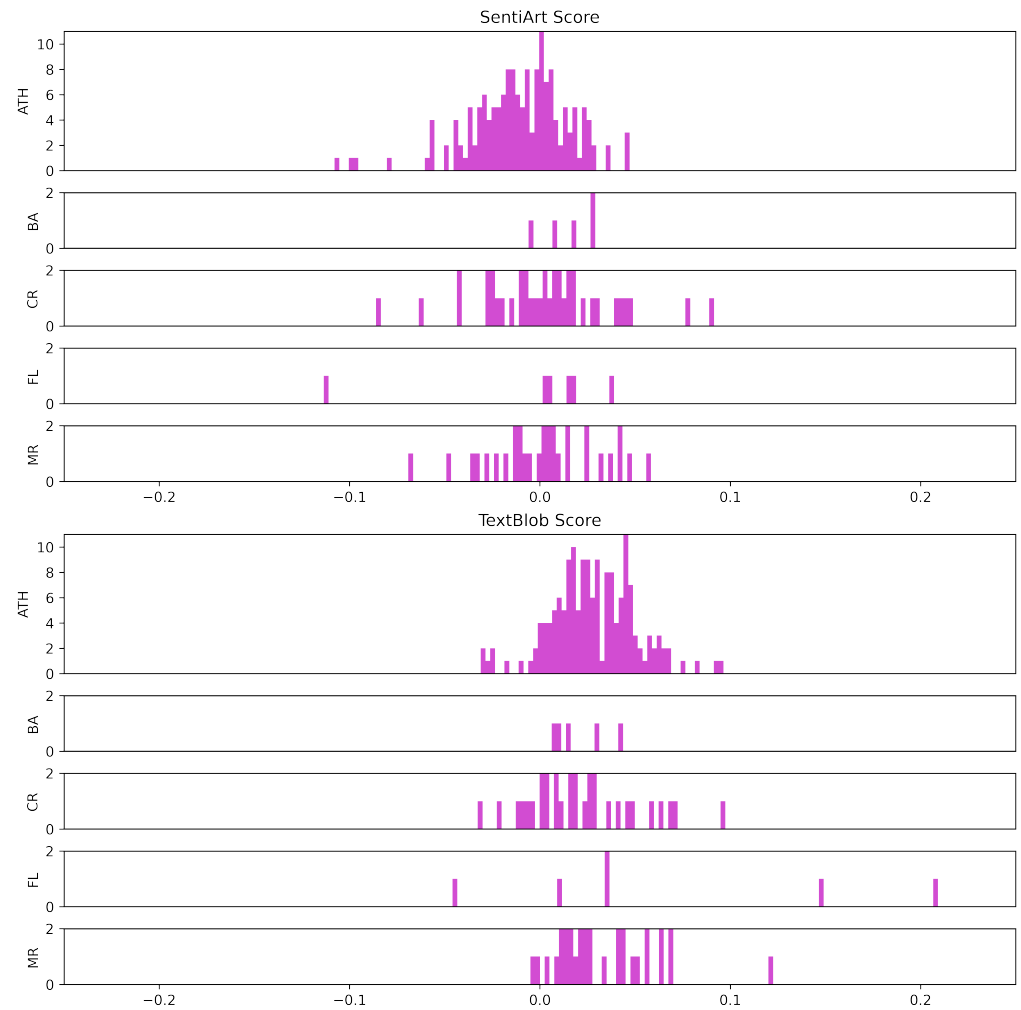


Figure 4: Distribution of sentiment scores across reviews in English journals.

Adapting an approach to sentiment analysis suggested by Jacobs (2019), we define the reference points by using what Jacobs calls “label words”. However, in contrast to Jacobs, who uses a theoretically and empirically tested set of emotion words, we use manually compiled lists of evaluative words that stood out as especially positive or negative in a close reading of a sample set of reviews.⁵

We then generated word2vec embeddings (Mikolov et al. 2013) for both languages, using the corpora and reviews as textual basis. For each manually determined label word, we added the words that were the most similar in the word embeddings⁶ to the respective lists of positive or negative label words. Then, we filtered both the label

5. Positive label words for English: *excellent, admirable, estimable, exemplary, invaluable, incomparable, superb, outstanding, wonderful, perfect, superior, worthy, fine, exceptional, skillful, masterful, extraordinary, impressive, notable, noteworthy*;

Negative label words for English: *terrible, grievous, hideous, ghastly, disgusting, unfavourable, disagreeable, distasteful, error, fault, unpleasant, imprudent, unlikely, undesirable, unreasonable, absurd, offensive, unsuitable, questionable, disconcerting*;

Positive label words for German: *anziehend, genial, geistreich, angemessen, wahr, poetisch, gelungen, ästhetisch, originell, künstlerisch, edel, großartig, dichterisch, meisterhaft, wertvoll, tadellos, wahrhaft, ideal, echt, hervorragend*;

Negative label words for German: *misform, überspannt, dürftig, seltsam, schädlich, unfertig, frech, enttäuschung, schwäche, tadel, simpel, übertrieben, überflüssig, fehler, niedrig, grauenhaft, umständlich, oberflächlich, mittelmäßig, unnatürlich*.

6. As the German word model is less stable, we only used the two most similar words, while for the English model, we were able to include the ten most similar words.

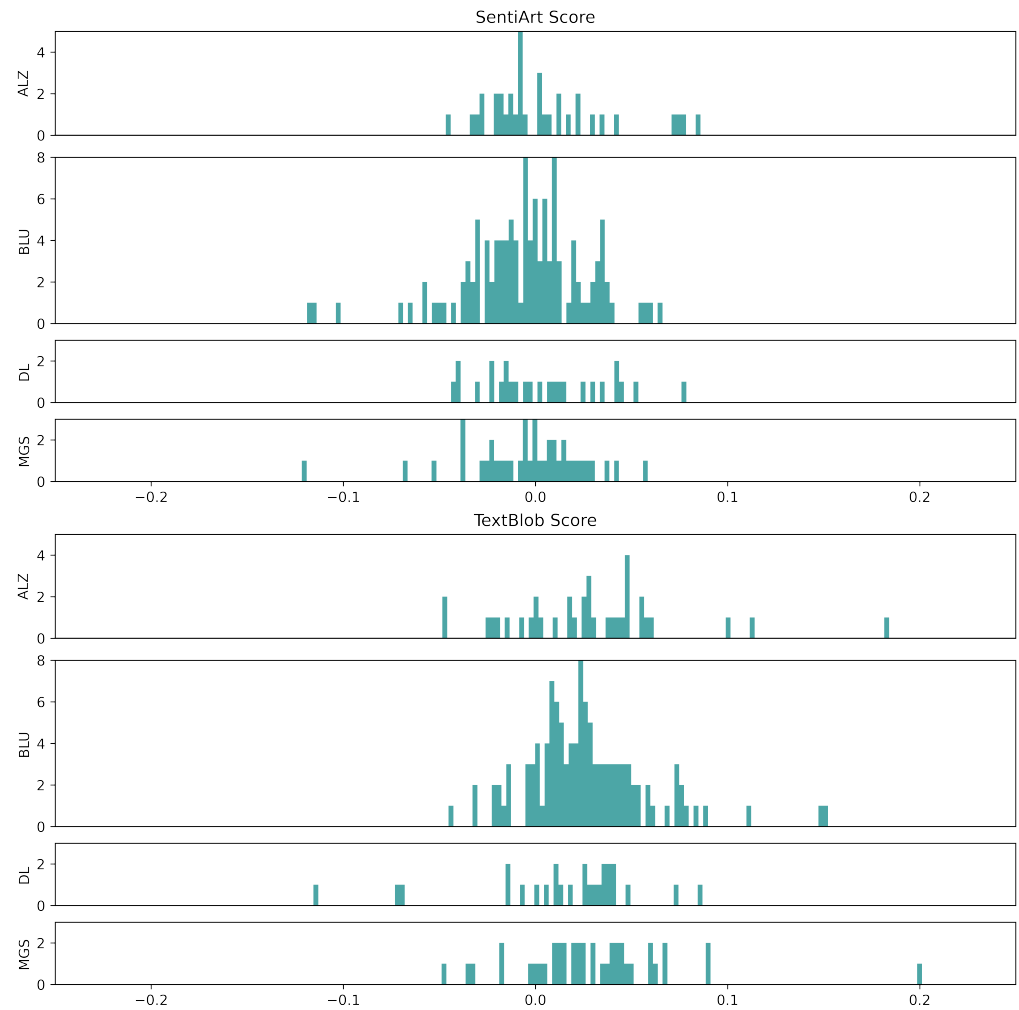


Figure 5: Distribution of sentiment scores across reviews in German journals.

words and newly added similar words according to the following criteria: With our approach, a focus on evaluative language on the word-level seems most practicable, which is why we excluded all word classes but adjectives and nouns. As an additional prerequisite, we only included words which appear more frequently in the reviews than in our corpora. By doing so, we model the particular register of reviews and thus exclude words predominantly used in the plot descriptions. Finally, to ensure some degree of generalizability, we only included words that belong to the 10,000 most frequent nouns and adjectives in all reviews.

After applying these limitations to the lists, we performed an affinity propagation clustering algorithm for both positive and negative evaluative words. This is necessary because all evaluative words are relatively close to each other in the word embeddings and combining positive and negative words helps to identify stable and unambiguous clusters. Then, we manually chose the most representative clusters to define the positive and negative poles of evaluation, represented by the centroid of each of these clusters. Based on the centroids, we calculated the cosine similarities between the positive and negative clusters and each word that belongs to the 10,000 most frequent adjectives or nouns more typically used in reviews. By subtracting the normalized sum of the negative similarities from the normalized sum of the positive similarities, we determined

whether a specific word is closer to the positive or negative cluster centroids.

Compared to TextBlob, our *ad hoc* SentiArt approach performs better at recognizing negative reviews: 12 out of 15 English and 10 out of 15 German negative reviews in the test set were attributed correctly. However, the SentiArt implementation performs worse for positive reviews, correctly identifying 9 positive English and 8 positive German reviews. The distributions across reviews from different journals in Figure 4 and 5 show that the SentiArt approach generally produces more negative scores, especially for English reviews. To make use of the strengths of both implementations, we conducted the analyses separately with the scores from the TextBlob and SentiArt approaches, as well as with a combination of both.

5.2 Text Features

Based on two publications surveying text features in stylistic and authorship attribution studies (Lagutina et al. 2019; Stamatatos 2009), we considered several textual levels for extracting features which are generally associated with a text's quality, complexity, and distinctiveness. The feature selection builds, however, also on more theory-based grounds since complexity and distinctiveness can be linked to a text's reception. A less complex text might have been written with a non-expert as the intended reader in mind, and might consequently appeal to a broader audience. A text's distance to other texts and position within the corpus can cause positive reactions (because the text is innovative and exciting) or negative reactions (because it exceeds the audience's horizon of expectation). An overview of all features is presented in Table 1.

Due to the limited size of our corpora, we split the texts into chunks of 200 sentences and calculated the features for each chunk, treating it as a separate document. This led to some loss of data since we excluded a text's last section if it was too short to constitute a full chunk. Not all features can be calculated for chunks; the semantic features (see Table 1) need to be calculated for a whole document because they are measures of the distance between chunks. If a feature's nature permitted that it was calculated for chunks (here called chunk-type features, as opposed to document-type features), we also calculated it for whole texts, treating the entire text as a single chunk. This way, we obtained two datasets: the chunk-based dataset, which contains the chunk-type features for each chunk, and the document-based dataset which contains the document-type features plus the chunk-type features calculated on the whole texts. We combined these two datasets in two ways: First, the document-based dataset was left unchanged and combined with the average across the chunks of a text in the chunk-based dataset. For the second dataset, the document-type features were copied and added to the chunk-based dataset of each respective text. We refer to the first combined dataset as the *document plus averaged chunks dataset* and to the second as the *chunks plus copied document dataset*.

For the feature extraction on the level of characters, we included the ratio of various special signs (punctuation marks, whitespaces, uppercase letters, commas, exclamation and question marks), while on the word-level, we used the ratio of unique uni-, bi-, and trigrams as well as the type-token-ratio as measures of lexical diversity, and the uni-,

Table 1: Text features.

		Chunk-type	Document-type
Basic text features	Character	Character frequency Ratio of punctuation marks Ratio of whitespace Ratio of exclamation marks Ratio of question marks Ratio of commas Ratio of uppercase letters	
	Lexical	Type-token ratio <i>n</i> -grams Ratio of unique unigrams Ratio of unique bigrams Ratio of unique trigrams Unigram entropy Bigram entropy Trigram entropy Corpus distance Unigram corpus distance Selective unigram corpus distance Bigram corpus distance Trigram corpus distance	
Distinctiveness	Semantic		Intra-textual variance Stepwise distance Outlier score Overlap score
Complexity	Syntactic	Tag distribution Production rule distribution Tag unigrams Tag bigrams Tag trigrams	
	Text Length	Average number of words per sentence Maximum number of words per sentence Average word length Average paragraph length Chunk text length	
	Other	Flesch reading ease score	

bi-, and trigram entropy.⁷

Established features in stylistic analyses such as tf-idf, bag-of-words representations, and n -gram frequencies (Lagutina et al. 2019) have the disadvantage that every word or n -gram constitutes an individual feature, leading to high-dimensional datasets on which classifiers easily overfit. As an alternative, we developed a measure called corpus distance, which is the cosine distance between a text's word frequency or n -gram frequency vector and the average word frequency or n -gram frequency vector of the rest of the corpus. We calculated the corpus distance for uni-, bi-, and trigrams. To account for named entities – as, for example, names of people or places that are unique to the story – an n -gram had to occur in at least two corpus texts to contribute to the distance. We also added a second version of the unigram corpus distance, where a word had to occur in at least 5 percent but no more than 50 percent of the documents, with the goal of finding words that are particular to selective writing styles.

To account for a text's semantic complexity, we used four measures introduced by Cranenburgh, Dalen-Oskam, and Zundert for computing different concepts of distance between the chunks of a text (2019). We calculated each of them with both document embeddings (Le and Mikolov 2014) and sentence BERT (SBERT) embeddings (Reimers and Gurevych 2019). Intra-textual variance measures how similar the individual chunks are to the average of all chunks making up a document, the centroid, while stepwise distance is a measure of the distance between successive chunks. The outlier and overlap scores look at the similarity to other works in the corpus. The former is the smallest distance between the centroid and another document's centroid, while the latter is the share of chunks belonging to other documents among the k chunks that are nearest to the centroid, with k being the number of chunks in the text.

We also included features on the syntactic text level. Using the Natural Language Processing (NLP) library spaCy for Python, we tagged the words in each text with their part-of-speech (POS) and counted the number of single tags as well as the number of two or three tags occurring subsequently, here called the tag bigrams and tag trigrams. "ADJ-NOUN-VERB", for example, is such a tag trigram, which means that an adjective followed by a noun which is in turn followed by a verb occurs in the text. Due to the number of possible combinations, we included only the frequency of the 100 most common tag n -grams. The production rule distribution served as another syntactic feature, but is available only for the English texts. A production rule is the pattern according to which one grammatical part of a sentence is followed by another part. We used NLTK, a different Python NLP library, and included the frequency of the 100 most common production rules.

The average word length, average and maximum number of words in a sentence, average length of a paragraph and text length of a chunk are measures for the general complexity of the text. Combining average word length measured in syllables and average sentence length in words, the Flesch reading ease score accounts for how challenging it is to read a text (Flesch 1948). As previous research has indicated that there is a negative

7. Entropy is a measure of the information content of a sequence of symbols (Baeza-Yates and Ribeiro-Neto 1999; Bentz et al. 2017). If a sequence consists mostly of one symbol, the sequence's information content is low. If the symbols making up the sequence are distributed uniformly, the entropy is highest. n -gram entropy is a measure of how uniformly a text's uni-, bi-, or trigrams are distributed.

correlation between readability and literary success (Ashok et al. 2013), we assume that the underlying complexity affects a text's perceived difficulty and consequently its popularity with broader audiences.

5.3 Prediction

To test if historical reception is dependent on text features, we trained three classifiers: (1) a classifier predicting whether a text had been reviewed, (2) a classifier determining if the review's sentiment was positive, neutral, or negative, and finally (3) a classifier predicting if a text had been featured in a library catalog. Then, we also tried to predict the review sentiment with a regression. To find the optimal combination of parameters, we ran a nested cross-validation for each of the four tasks, where the outer cross-validation evaluated the models selected by the inner cross-validation.

5.3.1 Model Selection

Since we had different options for models, model parameters, and features, we ran a grid search to find the combinations that achieved the highest performance for each of the four tasks (binary and multi-class classification, catalog classification, and regression). Due to the small size of the dataset, it was not possible to reserve a part of the data to evaluate the performance of the selected models. Testing models on data which they were not trained on is important because the model that performed best during training might be overfit to the data. Overfitting can only be assessed with an independent dataset.

Instead, we trained and tested the models by splitting the data into five folds and then running five separate inner cross-validations, each using a different fold for evaluation and the remaining four folds as the training data. This nested cross-validation allowed us to use all data for training the models while having independent test data. For each inner cross-validation, the data was again split into five folds, the models were trained on four of the folds and evaluated on the fifth fold, and finally the model with the best average performance across the five inner folds was selected. This model was then trained again using all data from the four folds of the outer cross-validation on which the inner cross-validation was run, and then evaluated on the fifth fold. The downside of this approach is that each of the five inner cross-validations selected a different model, and the final predicted values are a combination of the predictions of five different models.

To avoid overfitting to an author's writing style instead of learning the textual features that might be connected with a text's positive reception, all works written by an individual author were put into the same fold.

Besides the task-specific models and parameters further detailed below (see Section 5.3.2 and Section 5.3.3), we ran a separate nested cross-validations for the document-based dataset, the chunk-based dataset, and the two combinations of the two datasets (see Section 5.2). As an additional measure against overfitting, we tested whether the performance increased if we excluded either the tag distribution or the production rule distribution (which is only available for English) or both from the features since each of them amounted to 100 columns in the dataset.

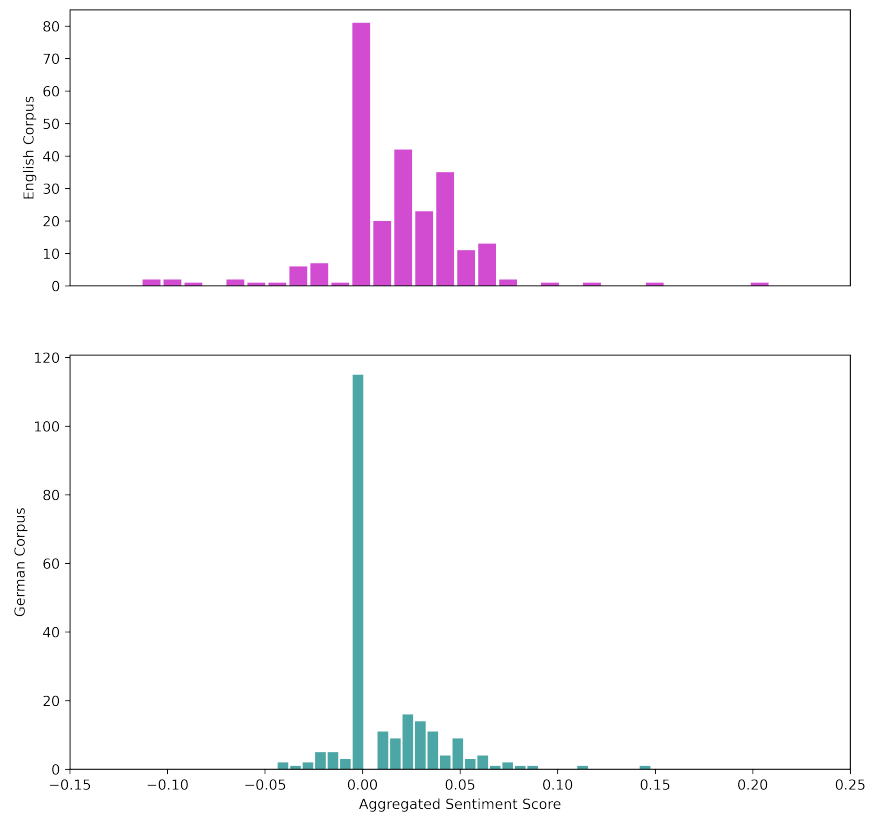


Figure 6: Distribution of aggregated sentiment scores for both corpora.

After running separate cross-validations for each feature combination, we selected the final best models by first finding the best models for each feature combination per fold of the outer cross-validation. Then, we chose the models associated with the feature level that had the highest mean validation score across all folds of the inner cross-validation.

5.3.2 Classification

As a first step, we tested if we could predict whether a text had been reviewed or not. The existence of a review is, as described in Section 4, the result of an evaluative selection decision by contemporaries, which means that even if the review was negative, the literary text generated enough attention to be reviewed. By the same logic, the inclusion in a circulating library catalog is also an indicator of interest by the broad public. Therefore, we used the same approach for classifying a text according to whether it had been included in a circulating library catalog. In the next step, we ran a classification with four classes to predict not only if a text had been reviewed, but also if the sentiment of the review(s) had been positive, negative, or neutral.

We tested the performance of two classifiers, SVM and XGBClassifier, and a selection of options for their respective parameters.⁸ SVMs are algorithms that fit a hyperplane which separates the data points belonging to different classes. XGBoost is a Python machine learning library that uses decision tree ensembles to make predictions.⁹ We only tested the document-based dataset and document plus averaged chunks dataset, since using chunk-level features would mean that the chunks making up a text could be placed into different classes. The results of a chunk-level classification would be even more difficult to interpret for multi-class classifications, since one would have to justify how severe the misclassifications into the different classes are relative to each other.

We used a combination of the scores from SentiArt and TextBlob, where only texts with clearly positive TextBlob-scores or clearly negative SentiArt-scores were labeled as either positive or negative and all others as neutral¹⁰ (see Figure 6). If a text had been reviewed multiple times, we aggregated the class assignments so that each text had only one label in the end. Texts that had both positive and negative reviews were excluded from the dataset, which was the case for six texts in the English corpus and for three texts in the German corpus. If a text had neutral and positive or neutral and negative reviews, we assigned the dominant label, and the more extreme one if both labels were equally frequent. We only included texts that were published between the year of the earliest review and the last year for which we surveyed literary magazines, so that texts that had no chance of being reviewed did not distort the classification. For the library catalog classification, we left out any text that was published after the last year for which we had searched the catalogs.

Due to the inclusion of the non-reviewed texts, the data contained approximately twice as many non-reviewed texts as reviewed texts. In addition, due to the exclusion of texts if they had contradicting reviews and the tendency of reviews to be positive, our data was heavily imbalanced for the multi-class classification and negatively reviewed texts were especially underrepresented. The number of reviewed texts in each class after filtering for publication years is shown in Table 2. The majority of English texts were included in a library catalog, which is why this dataset is also imbalanced (see Table 3, again filtered for publication year).

In both grid search and evaluation, we accounted for this class imbalance: We used a stratified cross-validation, meaning that each fold had approximately the same number of texts from each class (while maintaining that all works by an author were in the same fold), so that all classes were represented in both the training and test set. We used balanced accuracy, which is the mean accuracy of each class, thereby giving each class equal weight as the evaluation metric for the two binary classifications, and the F1 score, which is the harmonic mean between precision and recall as the evaluation metric for multi-class classification. Using a macro average, it gives equal weight to each class as well.

To see if our results were better than chance, we calculated several baselines. As the most basic baseline, we assigned the most frequent class in every case. For all other baselines,

8. SVM: C. XGBClassifier: max_depth, learning_rate, colsample_bytree.

9. XGBClassifier and XGBRegressor are part of its scikit-learn interface.

10. The thresholds for neutral labels were deduced from the data: For the English reviews, the lowest 12.5 percent of positive and negative scores were labeled as neutral. Because the German reviews are more clustered around 0, we used a lower threshold of 6.25 percent.

Table 2: Number of reviews.

	English	German
Not reviewed	365	330
Negative	15	10
Neutral	63	86
Positive	113	77

Table 3: Number of texts featured in library catalogs.

	English	German
Not Featured	93	181
Featured	456	302

we randomly drew from the class labels according to a certain probability. We used either equal probability or the class frequency as the probability of each class for the two binary classifications. Analogously, we assigned uniform probability and probability proportional to the class frequency for the multi-class baseline. Then, we also only considered the two most frequent classes and assigned probabilities proportional to their new frequency in this two-class setting. Finally, we left out the most underrepresented class and assigned the most frequent one a probability of 0.5 and the other ones of 0.25.

5.3.3 Regression

We ran separate regressions for the TextBlob- and SentiArt-generated scores. If a text had multiple reviews, we took the average over the sentiment scores of the individual reviews. Then, we ran another regression with a combination of the scores from the two tools. As described in the previous section (Section 5.3.2), the scores were split into classes to label reviews as positive, negative, or neutral. We created the combined score by taking the TextBlob-scores if they were positive enough for a review to be classified as positive, the SentiArt-scores if they were negative enough for a review to be classified as negative, and the average of the two if a review had been labeled as neutral.

We tested Support Vector Regression (SVR) and XGBRegressor as the regression models and combinations of their respective parameters in the grid search¹¹ along with the four feature levels. For evaluating the performance of the models, we calculated the correlation between true and predicted labels with Pearson's r , and its p -value. The p -value of each tested model was then calculated by taking the harmonic mean of the p -values across the folds of the respective inner cross-validation (Wilson 2019).

6. Results

6.1 Classification

As described in Section 5.3.1, we dealt with the problem of having only limited data by implementing a nested cross-validation choosing the feature combination that had the highest mean score across the folds of the inner cross-validation. For all the classification

11. SVR: C, epsilon. XGBRegressor: max_depth, learning_rate, colsample_bytree.

Table 4: Crosstab for reviewed/not reviewed classification, English.

		Predicted		
		Not reviewed	Reviewed	Total
True	Not reviewed	259	106	365
	Reviewed	26	165	191
	Total	285	271	556

Table 5: Crosstab for reviewed/not reviewed classification, German.

		Predicted		
		Not reviewed	Reviewed	Total
True	Not reviewed	227	103	330
	Reviewed	72	101	173
	Total	299	204	503

tasks, using only the document-based dataset was the best choice of features for both languages. The crosstabs (Table 4 to 9) show how many texts from each class were predicted to be a specific class.

6.1.1 Reviewed/Not Reviewed

The best models achieved a balanced accuracy of 0.787 (baseline = 0.531) for English and 0.636 (baseline = 0.539) for German (cf. Table 4 and 5).

6.1.2 Multi-class Classification

The F1 score of the best model for English is 0.406 (baseline = 0.261) and 0.342 (baseline = 0.244) for German (cf. Table 6 and 7).

6.1.3 Library Catalogs Classification

The balanced accuracy of the predicted labels for English is 0.617 (baseline = 0.510) and for German 0.539 (baseline = 0.517) (cf. Table 8 and 9).

Table 6: Crosstab for multi-class classification, English.

		Not reviewed	Predicted			Total
			Negative	Neutral	Positive	
True	Not reviewed	265	3	45	52	365
	Negative	6	3	1	5	15
	Neutral	15	2	23	23	63
	Positive	44	5	23	41	113
	Total	330	13	92	121	556

Table 7: Crosstab for multi-class classification, German

		Not reviewed	Predicted			Total
			Negative	Neutral	Positive	
True	Not reviewed	222	6	54	48	330
	Negative	6	0	2	2	10
	Neutral	38	0	27	21	86
	Positive	26	0	18	33	77
	Total	292	6	101	104	503

Table 8: Crosstab for library catalogs classification, English.

		Not featured	Predicted	
			Featured	Total
True	Not featured	53	40	93
	Featured	153	303	456
	Total	206	343	549

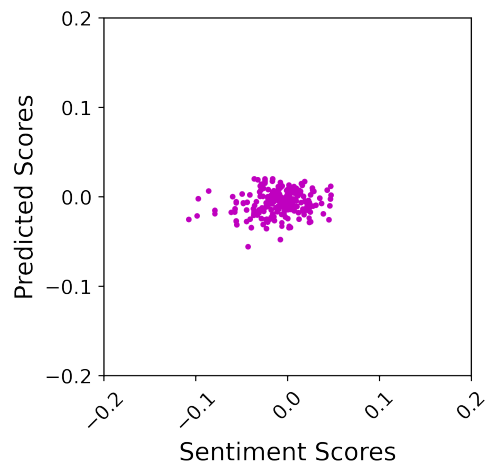
Table 9: Crosstab for library catalogs classification, German.

		Not featured	Predicted	
			Featured	Total
True	Not featured	91	90	181
	Featured	128	174	302
	Total	219	264	483

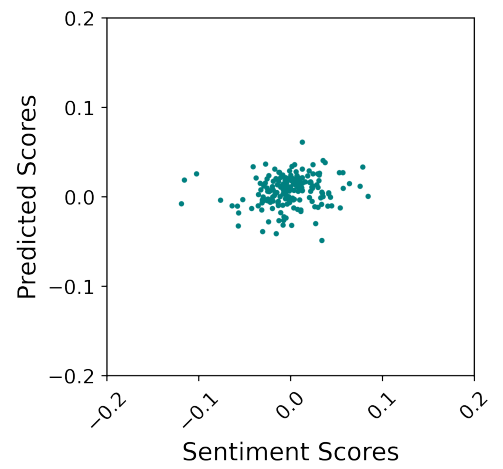
6.2 Regression

The highest significant correlation coefficient from the cross-validation, or the highest coefficient if none was significant, are reported in Table 10. In Figure 7, the true and the predicted scores are plotted against each other.

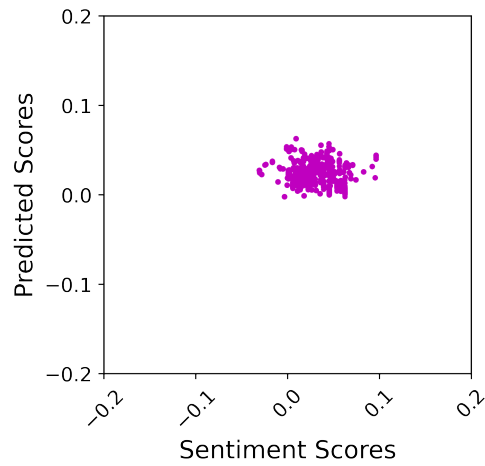
Unlike for the classification, there is no consistent best feature combination for all regression tasks. The chunks plus copied document dataset was the best choice for English on the SentiArt and TextBlob labels, and the document plus averaged chunks dataset for the combined labels. For German, the chunks plus copied document dataset had the best performance with TextBlob and combined labels, while the document-based dataset performed best with the SentiArt labels.



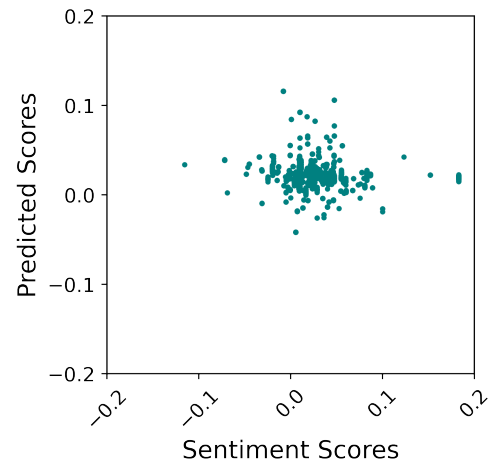
(a) English, SentiArt.



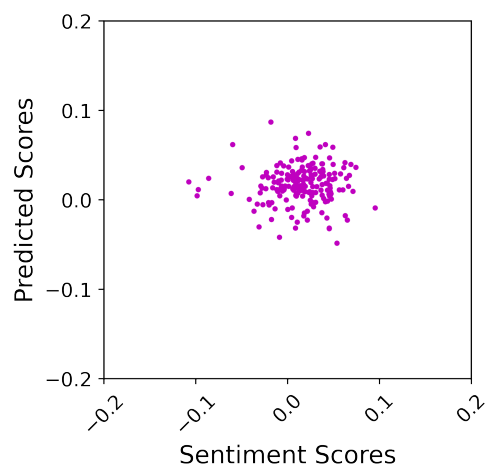
(b) German, SentiArt.



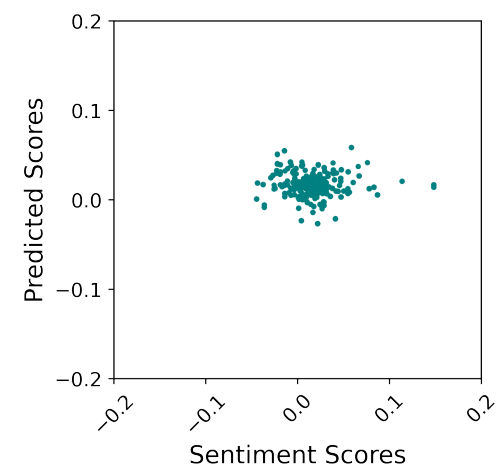
(c) English, TextBlob.



(d) German, TextBlob.



(e) English, Combined.



(f) German, Combined.

Figure 7: Sentiment scores and predicted scores.

Table 10: Regression results.

	English	German
SentiArt	0.113***	0.168**
TextBlob	-0.019	-0.111***
Combined	0.015	-0.126***

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

7. Discussion

In all experiments, our classifiers performed better than the baselines for the three classification tasks, but only barely so for the library catalog classification with texts in German. Overall, we were able to achieve better classification results with the English than the German dataset. For both languages, predicting whether a text had been reviewed generated accuracy values that exceed the threshold of better-than-chance. This can be seen as an indication that texts that generate enough interest to receive a review share certain textual qualities. By suggesting such a relationship, the results may be seen as a consolidation of the theory presented in previous research (see Heydebrand and Winko 1996, 99) that the existence of a review alone – may it be positive or negative – is an important structuring device representing the attention a text attracted. The fact that a text had been added to a circulating library might be viewed as a similar indicator of interest by a broad audience. At least for the English texts, our findings indicate that there are some detectable textual qualities that can be linked to a sparked interest among the general reading public.

The higher balanced accuracy score for English library catalogs might also be connected to the distribution of data described in Section 4.2: In contrast to the German dataset, there is a clear tendency for late 19th and early 20th century catalogs to include contemporary texts, which can be assumed to be stylistically more homogeneous. Moreover, the last two catalogs surveyed are from the same library, *Mudie's*, whose owner Charles Edward Mudie allegedly only advertised books that satisfied his personal moral and literary standards (Katz 2017, Roberts 2006). It seems plausible that these factors led to more quantitatively detectable similarities within the set of texts advertised in library catalogs and therefore to a higher balanced accuracy score.

In general, the prediction of in- or exclusions in circulating library catalogs is, with its balanced accuracy values of 0.617 and 0.539, less conclusive than the prediction of reviews. On a first glance, these differences in the classification success are surprising: The data distribution for both reviews and circulating library catalogs is similarly imbalanced; the diverging results might therefore be linked to habitual aspects of the two literary institutions. While circulating library catalogs reflect the multi-faceted and inconsistent interests of the public, the selective decisions that lead to a text being reviewed in a literary journal are more concise and uniform. In other words, the selective decisions of a smaller group of expert readers seem to be easier to predict than those based on the expected popularity with a more diverse group of lay readers.

The low correlation coefficients of the best model with 0.113 for English and 0.168 for German texts show that even with our adapted SentiArt approach, there is only a very

weak correlation between the measured sentiment in reviews and textual markers of the reviewed texts. Due to this weak effect, we did not further analyze the contribution of individual features. Reasons for why combining the scores generated by the two tools leads to low and non-significant correlation coefficients could be inadequately set thresholds for switching from one tool to the other or the usage of average values for neutral reviews.

The mean validation scores across the folds of the inner cross-validations are close to the evaluation scores on the test data of the outer folds for the classification tasks, while they are a lot higher than the evaluation scores for the regression tasks. The best regression models in the inner cross-validation do not generalize well on new data and seem to overfit to the training data.

8. Conclusion

Modeling historical reception requires a dataset that encodes literary contexts by combining texts with complementary information on how they were received by their contemporaries. By operationalizing the theoretical framework suggested by Heydebrand and Winko (1996), we were able to formalize a text's reception by experts, as well as its popularity with audiences. Differentiating these two levels of literary evaluation allows a more detailed analysis of historical reception and lays the ground work for future research on synchronic reading, diachronic canonization, and their interplay.

A conclusion of our findings is difficult due to the limited amount of available data. Historical literary data is scarce, and a larger dataset might have led to different results. Based on this data-rich literary history dataset, predicting review sentiment from texts alone proved to be successful only to a very limited extent. We had better success predicting whether literary works had been reviewed or not: There seem to be certain text qualities that make it more likely that a reviewer will pay attention and choose to review a text. To some degree, such a relationship between textual markers and selective decisions can also be detected when examining texts featured in circulating library catalogs. As circulating libraries cater to a more diverse group of readers, we assume that diverging interests represented in the catalogs could be responsible for the lower balanced accuracy values.

The low predictability of reception along with the limited predictability of a text's canonization status, which we have shown in a previous publication (Brottrager et al. 2021), indicate that the assessment of historical novels and narratives by contemporaries as well as by a modern expert audience are mainly not due to measurable textual qualities, but the result of a complex interplay of selection and interpretation processes which are influenced by both literary and non-literary factors (Rippl and Winko 2013; Winko 2002).

Generally, our *ad hoc* SentiArt approach has proven to be useful for the sentiment analysis of historical reviews because it was – in comparison to the established TextBlob tool – more adept at identifying the particularities of evaluative language in reviews, as, for example, the implicitness and vagueness of negative comments. We will work on fine-tuning the word embeddings to increase the accuracy of the approach in the

detection of positive reviews. Additionally, it seems reasonable to add a time component to our approach: Our corpora comprise texts from a time span of over 200 years. During this time, the market for and the status of literature changed dramatically, as did the expectations of different generations of audiences and literary experts. These historical shifts in readers' and reviewers' perspectives are not yet accounted for in our experiments, and we assume that all reviews express a certain sentiment in reaction to the same textual features. By extracting period-specific evaluation words and computing period-specific evaluation scores, we could account for these changes in perception, reception, and expectations.

9. Data Availability

Corpora, reviews, metadata, trained word embeddings, and sentiment scores can be accessed via <https://doi.org/10.6084/m9.figshare.19672410.v1>.

10. Software Availability

The scripts are available at https://github.com/sta-a/jcls_reception.

11. Acknowledgements

This work is part of “Relating the Unread. Network Models in Literary History”, a project supported by the German Research Foundation (DFG) through the priority programme SPP 2207 Computational Literary Studies (CLS). Special thanks to Joël Doat for his advice on the formal aspects of word embeddings.

12. Author Contributions

Judith Brottrager: Formal Analysis, Data curation, Writing – original draft

Annina Stahl: Formal Analysis, Writing – original draft

Arda Arslan: Formal Analysis

Ulrik Brandes: Supervision, Writing – review & editing

Thomas Weitin: Supervision, Writing – review & editing

References

- Algee-Hewitt, Mark, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, and Hannah Walser (2016). “Canon/Archive. Large-scale Dynamics in the Literary Field”. In: *Pamphlets of the Stanford Literary Lab* 11. <https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf> (visited on 10/25/2022).
- Algee-Hewitt, Mark and Mark McGurl (2015). “Between Canon and Corpus: Six Perspectives on 20th-Century Novels”. In: *Pamphlets of the Stanford Literary Lab* 8. [http://litlab.stanford.edu/LiteraryLabPamphlet8.pdf](https://litlab.stanford.edu/LiteraryLabPamphlet8.pdf) (visited on 10/25/2022).

- Alvarado, Rafael C. (2019). "Digital Humanities and the Great Project: Why We Should Operationalize Everything - and Study Those Who Are Doing So Now". In: *Debates in the Digital Humanities 2019*. Ed. by Matthew K. Gold and Lauren F. Klein. University of Minnesota Press, 75–82. [10.5749/j.ctvg251hk](https://doi.org/10.5749/j.ctvg251hk).
- Ashok, Vikas, Song Feng, and Yejin Choi (2013). "Success with Style: Using Writing Style to Predict the Success of Novels". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1753–1764.
- Baeza-Yates, Ricardo and Berthier Ribeiro-Neto (1999). *Modern Information Retrieval – The Concepts and Technologies Behind Search*. Addison Wesley.
- Bentz, Christian, Dimitrios Alikaniotis, Michael Cysouw, and Ramon Ferrer-i-Cancho (2017). "The Entropy of Words — Learnability and Expressivity across More than 1000 Languages". In: *Entropy* 19 (6). [10.3390/e19060275](https://doi.org/10.3390/e19060275).
- Bode, Katherine (2018). *A World of Fiction: Digital Collections and the Future of Literary History*. University of Michigan Press.
- Brottrager, Judith, Annina Stahl, and Arda Arslan (2021). "Predicting Canonization: Comparing Canonization Scores Based on Text-Extrinsic and -Intrinsic Features". In: *Proceedings of the Conference on Computational Humanities Research 2021*. CEUR Workshop Proceedings, 195–205. http://ceur-ws.org/Vol-2989/short_paper21.pdf (visited on 10/25/2022).
- City University London (2001). *The Athenaeum Projects*. <https://athenaeum.city.ac.uk/> (visited on 04/22/2022).
- Cranenburgh, Andreas van, Karina van Dalen-Oskam, and Joris van Zundert (2019). "Vector Space Explorations of Literary Language". In: *Language Resources and Evaluation* 53 (4), 625–650. [10.1007/s10579-018-09442-4](https://doi.org/10.1007/s10579-018-09442-4).
- Du, Keli and Katja Mellmann (2019). "Sentimentanalyse als Instrument literaturgeschichtlicher Rezeptionsforschung". In: *DARIAH-DE Working Papers* 32. <http://webdoc.sub.gwdg.de/pub/mon/dariah-de/dwp-2019-32.pdf> (visited on 10/25/2022).
- Flesch, Rudolph (1948). "A New Readability Yardstick". In: *The Journal of Applied Psychology* 32 (3), 221–233.
- Gamer, Michael (2000). *Romanticism and the Gothic: Genre, Reception, and Canon Formation*. 40. Cambridge University Press.
- Garside, Peter (2011). *British Fiction 1800–1829: A Database of Production, Circulation & Reception*. British Fiction 1800–1829. <http://www.british-fiction.cf.ac.uk/> (visited on 04/22/2022).
- Garside, Peter and Rainer Schöwerling, eds. (2000). *The English Novel 1770–1829 : A Bibliographical Survey of Prose Fiction Published in the British Isles*. Vol. 2: 1800–1829. Oxford University Press.
- GJZ18 (2021). *Gelehrte Journale und Zeitungen der Aufklärung*. GJZ18. <https://gelehrte-journale.de> (visited on 04/22/2022).
- Halliday, M. A. K. and Christian M. I. M. Matthiessen (2014). *Halliday's Introduction to Functional Grammar*. Routledge.
- Heydebrand, Renate von and Simone Winko (1996). *Einführung in die Wertung von Literatur*. Schöningh.
- Italia, Iona (2012). *The Rise of Literary Journalism in the Eighteenth Century: Anxious Employment*. Routledge.

- Jacobs, Arthur M. (2019). "Sentiment Analysis for Words and Fiction Characters From the Perspective of Computational (Neuro-)Poetics". In: *Frontiers in Robotics and AI* 6. [10.3389/frobt.2019.00053](https://doi.org/10.3389/frobt.2019.00053).
- Jacobs, Edward H. (2003). "Eighteenth-Century British Circulating Libraries and Cultural Book History". In: *Book History* 6 (1), 1–22. [10.1353/bh.2004.0010](https://doi.org/10.1353/bh.2004.0010).
- Jäger, Georg (1982). "Die Bestände deutscher Leihbibliotheken zwischen 1815 und 1860. Interpretation statistischer Befunde". In: *Buchhandel und Literatur: Festschrift für Herbert G. Göpfert zum 75. Geburtstag am 22. September 1982*. Harrassowitz, 247–313.
- Jockers, Matthew Lee (2013). *Macroanalysis: Digital Methods and Literary History*. Topics in the Digital Humanities. University of Illinois Press.
- Katz, Peter J. (2017). "Redefining the Republic of Letters: The Literary Public and Mudie's Circulating Library". In: *Journal of Victorian Culture* 22 (3), 399–417.
- Killer, Markus (2019). *textblob-de*. Version 0.4.4a1. <https://textblob-de.readthedocs.io/en/latest/index.html> (visited on 10/25/2022).
- Lagutina, Ksenia, Nadezhda Lagutina, Elena Boychuk, Inna Vorontsova, Elena Shliakhtina, Olga Belyaeva, Ilya Paramonov, and P. G. Demidov (2019). "A Survey on Stylometric Text Features". In: *25th Conference of Open Innovations Association (FRUCT)*, 184–195. [10.23919/FRUCT48121.2019.8981504](https://doi.org/10.23919/FRUCT48121.2019.8981504).
- Le, Quoc and Tomas Mikolov (2014). "Distributed Representations of Sentences and Documents". In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by Eric P. Xing and Tony Jebara. Beijing: PMLR, 1188–1196. <http://proceedings.mlr.press/v32/le14.pdf> (visited on 10/17/2022).
- Loria, Steven (2018). *textblob Documentation*. Release 0.15. <https://textblob.readthedocs.io/en/dev/index.html> (visited on 10/25/2022).
- Martin, J. R. and Peter Robert Rupert White (2007). *The Language of Evaluation: Appraisal in English*. Palgrave Macmillan.
- Martino, Alberto (1990). *Die deutsche Leihbibliothek: Geschichte einer literarischen Institution (1756-1914)*. Harrassowitz.
- Mikolov, Tomas, Kai Chen, Gregory S. Corrado, and Jeffrey Dean (2013). "Efficient Estimation of Word Representations in Vector Space". In: *arXiv preprint*. [10.48550/arXiv.1301.3781](https://arxiv.org/abs/1301.3781).
- Moretti, Franco (2013). *Distant Reading*. Verso.
- Pichler, Axel and Nils Reiter (2021). "Zur Operationalisierung literaturwissenschaftlicher Begriffe in der algorithmischen Textanalyse. Eine Annäherung über Norbert Althenhofers hermeneutische Modellinterpretation von Kleists *Das Erdbeben in Chili*". In: *Journal of Literary Theory* 15 (1), 1–29. [10.1515/jlt-2021-2008](https://doi.org/10.1515/jlt-2021-2008).
- Plachta, Bodo (2019). "Literaturzeitschriften". In: *Grundthemen der Literaturwissenschaft: literarische institutionen*. Ed. by Norbert Otto Eke and Stefan Elit. De Gruyter, 345–356.
- Porter, J.D. (2018). "Popularity/Prestige". In: *Pamphlets of the Stanford Literary Lab* 17. <https://litlab.stanford.edu/LiteraryLabPamphlet17.pdf> (visited on 10/25/2022).
- Raven, James and Antonia Forster, eds. (2000). *The English Novel 1770–1829 : A Bibliographical Survey of Prose Fiction Published in the British Isles*. Vol. 1: 1770-1779. Oxford University Press.
- Reimers, Nils and Iryna Gurevych (2019). "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language*

- Processing*. Association for Computational Linguistics, 3982–3992. [10.18653/v1/D19-1410](https://doi.org/10.18653/v1/D19-1410).
- Remus, Robert, Uwe Quasthoff, and Gerhard Heyer (2010). “SentiWS - A Publicly Available German-language Resource for Sentiment Analysis”. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/490_Paper.pdf (visited on 04/12/2022).
- Reul, Christian, Dennis Christ, Alexander Hartelt, Nico Ballbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttner, and Frank Puppe (2019). “OCR4all — An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings”. In: *Applied Sciences* 9 (22). [10.3390/app9224853](https://doi.org/10.3390/app9224853).
- Rippl, Gabriele and Simone Winko (2013). “Einleitung”. In: *Handbuch Kanon und Wertung: Theorien, Instanzen, Geschichte*. Ed. by Gabriele Rippl and Simone Winko. J.B.Metzler.
- Roberts, Lewis (2006). “Trafficking Literary Authority: Mudie’s Select Library and the Commodification of the Victorian Novel”. In: *Victorian Literature and Culture* 34 (1), 1–25. [10.1017/S1060150306051023](https://doi.org/10.1017/S1060150306051023).
- Schröter, Julian, Keli Du, Julia Dudar, Cora Rok, and Christof Schöch (2021). “From Keyness to Distinctiveness – Triangulation and Evaluation in Computational Literary Studies”. In: *Journal of Literary Theory* 15 (1), 81–108. [10.1515/jlt-2021-2011](https://doi.org/10.1515/jlt-2021-2011).
- Stamatatos, Efstathios (2009). “A Survey of Modern Authorship Attribution Methods”. In: *Journal of the American Society for Information Science and Technology* 60 (3), 538–556. [10.1002/asi.21001](https://doi.org/10.1002/asi.21001).
- Underwood, Ted (2019). *Distant Horizons: Digital Evidence and Literary Change*. The University of Chicago Press.
- Underwood, Ted and Jordan Sellers (2016). “The Longue Durée of Literary Prestige”. In: *Modern Language Quarterly* 77 (3). [10.1215/00267929-3570634](https://doi.org/10.1215/00267929-3570634).
- Wilson, Daniel (2019). “The Harmonic Mean p-Value for Combining Dependent Tests”. In: *Proceedings of the National Academy of Sciences of the United States of America* 116 (4), 1195–1200. [10.1073/pnas.1814092116](https://doi.org/10.1073/pnas.1814092116).
- Winko, Simone (2002). “Literatur-Kanon als invisible hand-Phänomen”. In: *Literarische Kanonbildung*. Ed. by Heinz Ludwig Arnold. edition text + kritik, 9–24.